

## *Bacillus subtilis* Genome Diversity<sup>∇†</sup>

Ashlee M. Earl,<sup>1</sup> Richard Losick,<sup>2</sup> and Roberto Kolter<sup>1\*</sup>

Department of Microbiology & Molecular Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115,<sup>1</sup> and Department of Molecular & Cellular Biology, The Biological Laboratories, Harvard University, 16 Divinity Ave., Cambridge, Massachusetts 02138<sup>2</sup>

Received 23 August 2006/Accepted 7 November 2006

**Microarray-based comparative genomic hybridization (M-CGH) is a powerful method for rapidly identifying regions of genome diversity among closely related organisms. We used M-CGH to examine the genome diversity of 17 strains belonging to the nonpathogenic species *Bacillus subtilis*. Our M-CGH results indicate that there is considerable genetic heterogeneity among members of this species; nearly one-third of *Bsu168*-specific genes exhibited variability, as measured by the microarray hybridization intensities. The variable loci include those encoding proteins involved in antibiotic production, cell wall synthesis, sporulation, and germination. The diversity in these genes may reflect this organism's ability to survive in diverse natural settings.**

Whole-genome sequence comparisons of different bacteria have led to many surprising observations. Prime among these is the remarkable genomic variability displayed by some phylogenetically cohesive units that we call bacterial species (2). In some cases isolates that exhibit 100% sequence identity at the level of 16S rRNA exhibit as little as 40% conservation in total gene content (39). Yet other species appear to have remarkably conserved genomes (13, 31, 32, 40). The ecological and evolutionary significance of the diversity of genome structures within different species remains very much unexplored. Also, because the majority of intraspecies comparisons carried out thus far have involved pathogens, our current view of genome evolution is limited. Here we analyzed the genome structure and diversity of *Bacillus subtilis*, a nonpathogenic, spore-forming bacterium commonly found in soil.

*B. subtilis* is arguably one of the best known and most extensively studied gram-positive bacteria (30). While a great deal is known about *B. subtilis* at the molecular level, relatively little is known about its ecology and evolution. An analysis of restriction fragment length polymorphisms in three loci (*rpoB*, *polC*, and *gyrA*) showed that there was considerable diversity in a collection of *B. subtilis* isolates obtained from desert soils (27). This analysis revealed that the isolates examined formed two distinct phylogenetic groups, prompting the proposal that two subspecies should be recognized (25). Despite the fact that all of these strains exhibited  $\geq 99.8\%$  sequence identity in their 16S rRNA genes, DNA-DNA reassociation analyses revealed a level of intersubspecies DNA relatedness of 58 to 69% (25). The level of intrasubspecies DNA relatedness was also shown to be as low as 82% (25), making it clear that even within a subspecies there is significant genome diversity. The following questions thus remain. How different are genomes of *B. subtilis* isolates? Which genes contribute to these differences? We

have begun to address these questions by examining diverse *B. subtilis* strains using sequence analysis of the conserved *gyrA* gene and comparative genomic hybridization, a microarray-based technique for whole-genome comparison.

To explore genome diversity in *B. subtilis* in more detail, we first collected several strains of the two subspecies, *B. subtilis* subsp. *subtilis* and *B. subtilis* subsp. *spizizenii* (Table 1). We specifically chose isolates from diverse geographic locations, including nondesert locales, based on the expectation that they should not be recent descendants of one another. We also examined two *B. subtilis* strains that are thought to be closely related to the sequenced strain *Bsu168* (BS5 and the "wild" Marburg strain NCIB3610), as well as the type strain of *Bacillus vallismortis* (DV1-F-3), the closest known relative of *B. subtilis* (28).

We first assessed strain relatedness by examining nucleotide variation at the highly conserved *gyrA* locus using previously described primers for PCR amplification of the gene (27). The phylogenetic tree in Fig. 1 is based on an alignment of 754-bp fragments amplified from an internal region of the *gyrA* gene. This tree shows the relationships among the strains listed in Table 1 when the *gyrA* gene from *Escherichia coli* K-12 was used as an outgroup. As expected, *B. vallismortis* DV1-F-3 is phylogenetically divergent from both *E. coli* K-12 and all *B. subtilis* strains. Also supporting the results of previous restriction fragment length polymorphism analyses (27), all of the *B. subtilis* isolates fell into one of two bootstrap-supported clusters. A similar analysis using 16S rRNA gene sequences from each strain failed to distinguish the *B. subtilis* subspecies or *B. vallismortis* as phylogenetically distinct taxa due to the limited number of informative sites at these loci (data not shown). Based on the *gyrA* analysis, we were able to assign 12 of our strains to *B. subtilis* subsp. *subtilis* and the remaining 6 strains to *B. subtilis* subsp. *spizizenii*. We also examined three other conserved loci, *recA*, *pycA*, and *pyrG*, in a subset of our strains (data not shown). We found that *gyrA* proved to be as reliable a predictor of strain relatedness as any of the other markers. Interestingly, for all markers, the lengths of the branches within the *B. subtilis* subsp. *spizizenii* group suggest that there is greater genetic variability among members of this subspecies

\* Corresponding author. Mailing address: Department of Microbiology & Molecular Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115. Phone: (617) 432-1776. Fax: (617) 738-7664. E-mail: rkolter@hms.harvard.edu.

† Supplemental material for this article may be found at <http://jbb.asm.org/>.

∇ Published ahead of print on 17 November 2006.

TABLE 1. Strain list

Strain	Description, isolation site, and/or source	Reference(s)
<i>B. subtilis</i> subsp. <i>subtilis</i>		
Bsu168 (from France)	<i>trpC2</i>	6
Bsu168 <i>eps::tet</i>	168 derivative containing a 15-gene deletion of the <i>eps</i> operon	3
BS5	Tetracycline-sensitive strain derived from CU2189	7, 38
NCIB3610	Marburg, Germany "wild" strain; BGSC	4
1431	Moscow, Russia; BGSC	22
1440	Moscow, Russia; BGSC	22
HIKEMIN-7	Kee, Hawaii	This study
HLR11	Hyallite Reservoir, Montana	This study
MGl47	Mendenhall Glacier, Alaska	This study
Natto (= IFO3335)	Japan; BGSC	35
RO-FF-1 (= NRRL B-23074)	Mojave Desert, Rosamond, CA; USDA	27
RO-NN-1 (= NRRL B-14823)	Mojave Desert, Rosamond, CA; USDA	27
T-89-48 (= NRRL B-23076)	Sonoran Desert, Tumamoc Hill, AZ; USDA	19
<i>B. subtilis</i> subsp. <i>spizizenii</i>		
DV1-B-1 (= NRRL B-23054)	Death Valley National Monument, California; USDA	27
DV1-E-2 (= NRRL B-23050)	Death Valley National Monument, California; USDA	27
N10	Mediterranean Sea, coast of Egypt; BGSC	12
RO-E-2 (= NRRL B-23055)	Mojave Desert, Rosamond, CA; USDA	27
TU-B-10 (= NRRL B-23049)	Sahara Desert, Nefta, Egypt; USDA	27
W23 (= NRRL B-14472)	USDA	37
<i>B. vallismortis</i> DV1-F-3 (= NRRL B-14890)	Death Valley National Monument, California; USDA	27

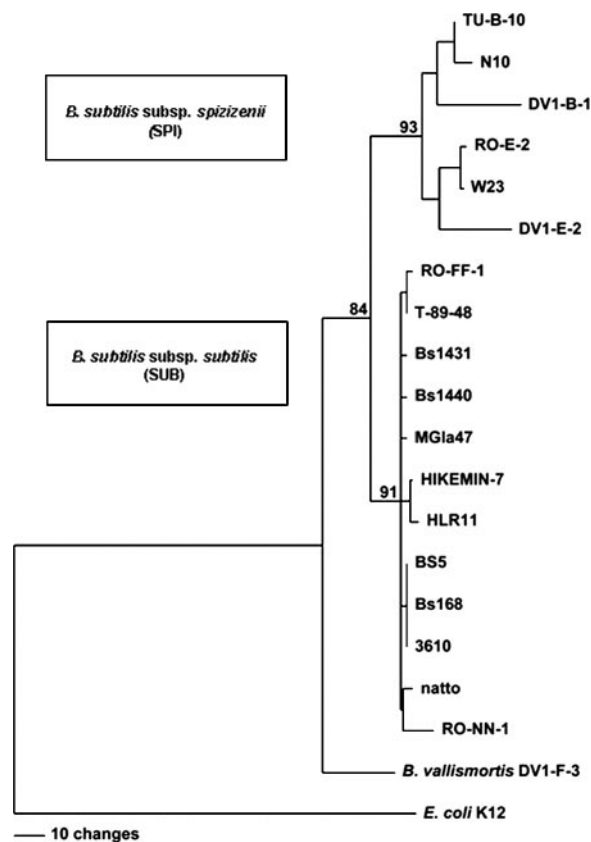


FIG. 1. Maximum parsimony tree derived using CLUSTALX and PAUP analysis of 754-bp *gyrA* sequences (34, 36). The *E. coli* K-12 outgroup is forced. The numbers at nodes indicate bootstrap support values, as calculated by PAUP.

than there is among members of *B. subtilis* subsp. *subtilis* (only data for *gyrA* are shown). While nucleotide information can reveal much about overall diversity and the relatedness among strains within a species, we were particularly interested in how the nucleotide variation translated to variation at the whole-genome level. We used microarray-based comparative genomic hybridization (M-CGH) to explore genomic diversity and to identify regions of genetic variability among members of both *B. subtilis* subspecies.

M-CGH has been used previously in analyses of a number of bacterial species (9, 20). Briefly, this technique allows one to predict gene absence (or divergence) versus gene presence by measuring the relative hybridization efficiencies of two differentially Cy-labeled pools of genomic DNA taken from two strains. We used Bsu168-specific oligonucleotide microarrays to identify genes that are absent or divergent in the strains listed in Table 1 compared to strain Bsu168, the only fully sequenced representative of the species. Each array was spotted with 3,722 gene-specific 60- to 70-mer oligonucleotides (designed by Compugen, San Jose, CA, and manufactured by SIGMA Genosys, The Woodlands, TX), representing ~91% of Bsu168's predicted gene set. Genes not represented in these experiments are located primarily in a large ~100-kb prophage, SP $\beta$ , as well as several other prophage regions located around the Bsu168 chromosome. Two micrograms of purified, Sau3AI-digested genomic DNA was labeled with either Cy3- or Cy5-conjugated dCTP as described previously (10). Following hybridization and scanning (8), the microarray images were loaded into Genepix 4.0 (Axon Instruments, Union City, CA) to calculate the ratios of Cy5 fluorescence intensity to Cy3 fluorescence intensity for all gene spots. Within a given experiment, gene spots with fluorescence intensities that were less than the average of the negative control spot intensities were excluded from the analysis. Ratios greater than this cutoff were then transformed by  $\log_2$ . Normalization to compensate for

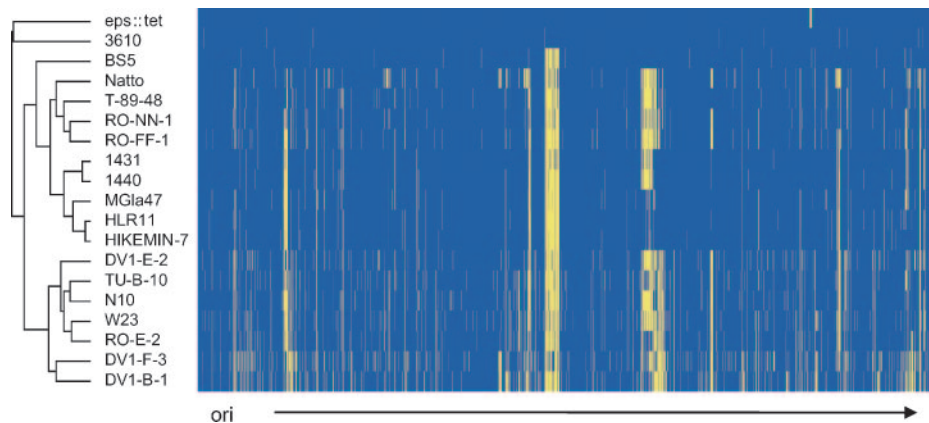


FIG. 2. M-CGH composite view of genome diversity exhibited by 18 *B. subtilis* strains and one *B. vallismortis* strain. Each row shows the results for one strain, and each column represents genes as they are found along the length of the Bsu168 genome, starting at the origin of replication and proceeding clockwise (*dnaA* to *rpmH*). Blue indicates a gene that is likely present in the test strain (average Cy control/Cy test  $\log_2$  ratio,  $\leq 1$ ), and yellow indicates a gene that is either absent or too highly divergent to hybridize with equal efficiency to the gene of Bsu168 (average Cy control/Cy test  $\log_2$  ratio,  $>1$ ). The image was generated using the CLUSTER and TREEVIEW programs (11).

slide-to-slide variation was accomplished by adjusting all ratios so that the median of all  $\log_2$  ratios for a given experiment equaled zero. The comparative hybridizations were repeated for each test strain three or four times and included at least one hybridization where the labeling regimen was switched to rule out potential bias introduced by inherent differences in Cy dye incorporation. The final data sets represent median values of these results (see Table S1 in the supplemental material). A gene was considered absent or divergent if the  $\log_2$  ratio of Bsu168 fluorescence to test strain fluorescence at a given gene spot was greater than 1. As controls, Bsu168-Bsu168 and Bsu168-Bsu168 *eps::tet* hybridizations were performed (Table 1). As expected, the self-self control experiments yielded no genes with a  $\log_2$  fluorescence ratio greater than 1. The results of the Bsu168-Bsu168 *eps::tet* hybridization did, however, reveal a potential limitation of the array; the values for only 14 of the 15 genes known to be deleted in this strain were above the cutoff ratio for gene absence or divergence. We believe that this may have been a consequence of cross-hybridization between gene spots, although we have not formally tested this hypothesis.

Figure 2 is a composite view of the results of the M-CGH experiments in heat map format and organized according to the results of hierarchical clustering using Spearman rank correlation to calculate similarity among all data sets (11). It is immediately clear from this visual representation of the data that *B. subtilis*, as a species, exhibits a fairly high level of genomic variability. For the strains examined, a range in the amount of gene absence or divergence was detected by the array (Fig. 3). Mirroring what was obtained in the *gyrA* analysis, members of *B. subtilis* subsp. *subtilis* exhibited less diversity relative to Bsu168 (2.0 to 8.9% divergence among the open reading frames tested) than members of *B. subtilis* subsp. *spizizenii* (10.5 to 16.6% divergence among the open reading frames tested). The *B. vallismortis* representative, DV1-F-3, exhibited greater gene diversity (17% divergence among the open reading frames tested) than all of the *B. subtilis* subsp. *spizizenii* members except DV1-B-1. While DV1-B-1 exhibited slightly less gene diversity than DV1-F-3, it was strikingly more

diverse (6% greater gene diversity relative to Bsu168) than the other representative *B. subtilis* subsp. *spizizenii* members. Interestingly, the phylogeny obtained when *gyrA* was used as a marker of relatedness was in almost perfect concordance with the phylogeny obtained when the degree and pattern of gene variation measured by the arrays were considered. As reported recently for other microbial species (2, 15), M-CGH may also prove to be a reliable phylogenetic tool for typing strains of *B. subtilis*.

We have recently used NCIB3610 as an undomesticated *B. subtilis* strain because it forms highly structured multicellular communities (4, 5). Our microarray results suggest that of all the strains that we have analyzed, this “wild” Marburg strain is the closest relative of Bsu168. Bsu168 and NCIB3610 have identical *gyrA* sequences and exhibit no significant diversity at any of the genes represented on the microarray. In addition, a direct comparative sequence analysis of Bsu168’s genome and contigs taken

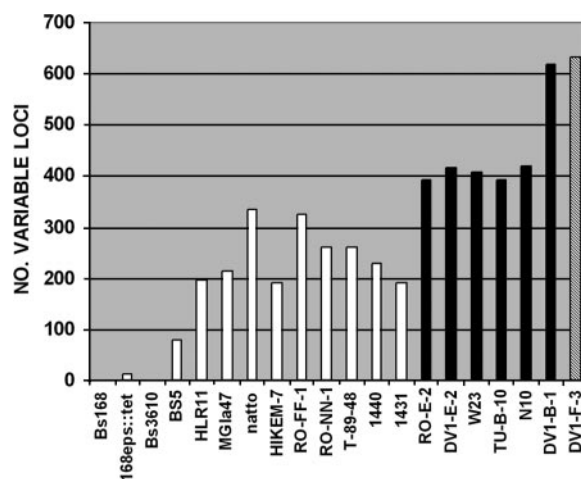


FIG. 3. Number of “variable” loci exhibited by each strain as determined by M-CGH. The white and black bars show data for *B. subtilis* subsp. *subtilis* and *B. subtilis* subsp. *spizizenii* strains, respectively, and the hatched bar shows data for *B. vallismortis*.

TABLE 2. Category role breakdown of divergent loci<sup>a</sup>

Category role	Total no. of genes in category		% of total genes divergent in category
	Bsu168	Divergent <sup>b</sup>	
<b>1.1. Cell wall</b>	<b>89</b>	<b>24 (11)<sup>c</sup></b>	<b>27</b>
1.10. Transformation/competence	25	2	8
<b>1.2. Transport/binding proteins and lipoproteins</b>	<b>400</b>	<b>123 (20)</b>	<b>31</b>
<b>1.3. Sensors (signal transduction)</b>	<b>39</b>	<b>14 (1)</b>	<b>36</b>
1.4. Membrane bioenergetics (electron transport chain and ATP synthase)	82	5 (2)	6
1.5. Mobility and chemotaxis	55	7 (2)	13
1.6. Protein secretion	26	6 (4)	23
1.7. Cell division	22	0	0
1.8. Sporulation	164	28 (9)	17
<b>1.9. Germination</b>	<b>26</b>	<b>11</b>	<b>42</b>
<b>2.1.1. Specific pathways</b>	<b>226</b>	<b>63 (14)</b>	<b>28</b>
2.1.2. Main glycolytic pathways	26	4	15
2.1.3. Tricarboxylic acid cycle	19	3	16
2.2. Metabolism of amino acids and related molecules	201	32 (5)	16
2.3. Metabolism of nucleotides and nucleic acids	92	20 (15)	22
2.4. Metabolism of lipids	89	13 (2)	15
2.5. Metabolism of coenzymes and prosthetic groups	103	12 (1)	12
2.6. Metabolism of phosphate	10	1	10
<b>2.7. Metabolism of sulfur</b>	<b>8</b>	<b>5</b>	<b>63</b>
3.1. DNA replication	26	3 (3)	12
<b>3.2. DNA restriction/modification and repair</b>	<b>42</b>	<b>12 (9)</b>	<b>29</b>
3.3. DNA recombination	19	1 (1)	5
3.4. DNA packaging and segregation	11	1 (1)	9
3.5.1. RNA synthesis—initiation	20	2 (1)	10
<b>3.5.2. RNA synthesis—regulation</b>	<b>224</b>	<b>82 (21)</b>	<b>37</b>
3.5.2. RNA synthesis—elongation	6	0	0
3.5.2. RNA synthesis—termination	4	0	0
3.6. RNA modification	28	6	21
3.7.1. Protein synthesis—ribosomal proteins	58	2	3
3.7.2. Protein synthesis—aminoacyl-tRNA synthetases	28	0	0
3.7.3. Protein synthesis—initiation	6	0	0
3.7.4. Protein synthesis—elongation	6	0	0
3.7.5. Protein synthesis—termination	3	0	0
3.8. Protein modification	35	6 (1)	17
<b>3.9. Protein folding</b>	<b>12</b>	<b>6 (2)</b>	<b>50</b>
4.1. Adaptation to atypical conditions	81	17 (1)	21
<b>4.2. Detoxification</b>	<b>89</b>	<b>37 (7)</b>	<b>42</b>
<b>4.3. Antibiotic production</b>	<b>35</b>	<b>23 (1)</b>	<b>66</b>
<b>4.4. Phage-related functions</b>	<b>87</b>	<b>54 (53)</b>	<b>62</b>
<b>4.5. Transposon and IS</b>	<b>10</b>	<b>8 (8)</b>	<b>80</b>
4.6. Miscellaneous	30	3 (1)	10
<b>5.1. Unknown (from <i>B. subtilis</i>)</b>	<b>540</b>	<b>152 (41)</b>	<b>28</b>
5.2. Unknown (from other organisms)	354	78 (17)	22
<b>6. Unknown (no similarity)</b>	<b>266</b>	<b>137 (72)</b>	<b>52</b>
Total	3,722	1,003 (326)	27

<sup>a</sup> Bold type indicates categories that have a higher number of divergent genes than the average total divergence for all genes in all functional categories.

<sup>b</sup> Only among *B. subtilis* strains.

<sup>c</sup> The numbers in parentheses indicate the number of divergent genes in the category that are found on phage-related elements or are predicted to be horizontally transferred (16).

from low-coverage (1×) sequencing of the NCIB3610 genome showed that there were virtually identical sequences at all the loci inspected (S. S. Branda, R. Kucherlapati, and R. Kolter, unpublished observation). There is, however, one notable difference between the two strains: NCIB3610 harbors an ~85-kb plasmid not found in Bsu168. Bsu168 was isolated in a screen for *B. subtilis*

biochemical mutants by Burkholder and Giles in 1947 (6), and since that time the identity of Bsu168's parent has been debated (17). While some workers believe that the "Marburg strain" used in the 1947 study is NCIB3610, this has not been conclusively shown. It is our opinion that based on the greater levels of diversity that are exhibited by other members of the *B. subtilis* subsp.

TABLE 3. Sporulation-related genes displaying variability

Gene		Category role	Predicted horizontal gene transfer	No. of strains (strain[s])
Bsu no.	Designation			
Bsu0023	<i>bofA</i>	Inhibition of pro-sigma-K processing machinery	No	1 (DV1-E-2)
Bsu0501	<i>rapI<sup>a</sup></i>	Response regulator aspartate phosphatase	Yes	12
Bsu0502	<i>phrI<sup>a</sup></i>	Phosphatase (RapI) regulator	Yes	13
Bsu0555	<i>cotP</i>	Probable spore coat protein	No	1 (DV1-B-1)
Bsu0571	<i>yddD</i>	Unknown; similar to unknown proteins from <i>B. subtilis</i>	No	3 (HIKEM-7, HLR11, MGla47)
Bsu1177	<i>cotX</i>	Spore coat protein (insoluble fraction)	No	1 (Natto)
Bsu1178	<i>cotW</i>	Spore coat protein (insoluble fraction)	No	1 (DV1-B-1)
Bsu1179	<i>cotV</i>	Spore coat protein (insoluble fraction)	No	4 (HIKEM-7, HLR11, DV1-B-1, Natto)
Bsu1427	<i>yknT</i>	Unknown; similar to sporulation protein sigma-E controlled	No	1 (RO-E-2)
Bsu1890	<i>rapK</i>	Response regulator aspartate phosphatase	Yes	12
Bsu1891	<i>phrK</i>	Phosphatase (RapK) regulator	Yes	12
Bsu1994	<i>sspC</i>	Small acid-soluble spore protein (minor alpha/beta-type small acid-soluble protein)	Yes	16
Bsu2338	<i>spoVAF</i>	Mutants lead to production of immature spores	No	6
Bsu2339	<i>spoVAE</i>	Mutants lead to production of immature spores	No	1 (DV1-B-1)
Bsu2352	<i>spoIIM</i>	Required for dissolution of the septal cell wall	No	1 (RO-FF-1)
Bsu2481	<i>yqgT</i>	Unknown; similar to gamma-D-glutamyl-L-diamino acid endopeptidase I	No	4 (DV1-E-1, DV1-B-1, TU-B-10, RO-NN-1)
Bsu2572	<i>spoIVCA</i>	Site-specific DNA recombinase required for creating the <i>sigK</i> gene (excision of the skin element)	Yes	3 (RO-FF-1, Natto)
Bsu2578	<i>rapE</i>	Response regulator aspartate phosphatase	Yes	6
Bsu2579	<i>phrE</i>	Phosphatase (RapE) regulator	Yes	1 (Natto)
Bsu2694	<i>yraD</i>	Unknown; similar to spore coat protein	No	6
Bsu3085	<i>cotS</i>	Spore coat protein	No	1 (DV1-B-1)
Bsu3086	<i>cotSA</i>	Spore coat protein	No	1 (DV1-B-1)
Bsu3087	<i>ytaA</i>	Unknown; similar to spore coat protein	No	1 (DV1-B-1)
Bsu3122	<i>tgl</i>	Transglutaminase	No	1 (DV1-E-2)
Bsu3603	<i>cotB</i>	Spore coat protein (outer)	No	5
Bsu3744	<i>phrF</i>	Phosphatase (RapF) regulator	No	2 (N10, DV1-E-2)
Bsu4096	<i>yjaA</i>	Unknown; similar to DNA-binding protein Spo0J-like	No	1 (BS5)

<sup>a</sup> Gene found on ICEBsI (a mobile genetic element), important for regulating excision and transfer of this element, and not thought to regulate sporulation (1).

*subtilis* group, coupled with the high levels of nucleotide identity determined by both the array and direct sequence comparisons, NCIB3610 and Bsu168 are directly related; NCIB3610 likely represents the progenitor of Bsu168.

All other strains examined exhibited much greater diversity relative to Bsu168. In fact, when all of the *B. subtilis* strains were considered together, we found that as many as 28% of the 3,722 genes represented on the microarray exhibited variability. Table 2 provides a tally of all the genes analyzed, grouped by their assigned category roles (<http://genolist.pasteur.fr/Subtilist/index.html>). Included in this table are the combined number of divergent loci and the level of total divergence for each specific category in only *B. subtilis* strains (DV1-F-3 was not included in the analysis). As predicted, there were category role groups (CRGs) that exhibited only limited or no variability, suggesting that genes in these categories are highly conserved among all *B. subtilis* isolates (e.g., category role 1.7 [cell division] and category role 3.7 [protein synthesis]). In contrast, 15 of the 44 CRGs exhibited a higher-than-average number of divergent loci; i.e., the number of genes displaying divergence for the group was higher than the average number of divergent genes for all CRGs. Not surprisingly, CRGs related to mobile genetic elements (category roles 4.4 and 4.5) were among this group. Also in agreement with previous comparative genomic

studies of other bacterial species, CRGs encoding proteins associated with the surface of the cell, including proteins involved in sensing and responding to the environment, were more divergent than other groups (24, 29, 42). This includes CRGs encoding proteins involved in germination, suggesting that the environmental cues recognized and/or the mechanism for exit from the dormant state may not be universal in this species.

Certainly one of the best-studied processes in *B. subtilis* is spore development, a programmed series of events that culminate in the formation of a highly resistant dormant cell (30, 33). While most sporulation-related genes identified in Bsu168 appear to be highly conserved, there are a few notable exceptions. Table 3 lists the sporulation-related genes that appear to be divergent in this collection of strains. Consistent with the idea that there is greater variability among proteins that physically interact and/or respond to the environment, more than one-half of the "divergent" sporulation-related genes encode spore coat constituents or encode proteins that function as part of the environmentally controlled phosphorelay system that ultimately governs when the cell enters sporulation (14, 30). Despite the observed diversity in these sporulation-related genes, all of the strains are able to form spores under laboratory conditions, suggesting that, like germination, some aspects

TABLE 4. "Essential" genes displaying variability

Gene		Category role	Predicted horizontal gene transfer	No. of strains (strain[s])
Bsu no.	Designation			
Bsu1522	<i>murD</i>	1.1. Cell wall		2 (N10, DV1-B-1)
Bsu2504	<i>yqfY</i>	1.1. Cell wall		1 (RO-NN-1)
Bsu3569	<i>tagH</i>	1.1. Cell wall	Yes	1 (Natto)
Bsu3569	<i>tagG</i>	1.1. Cell wall	Yes	6
Bsu3570	<i>tagF</i>	1.1. Cell wall	Yes	13
Bsu3572	<i>tagD</i>	1.1. Cell wall	Yes	11
Bsu3573	<i>tagA</i>	1.1. Cell wall	Yes	13
Bsu3574	<i>tagB</i>	1.1. Cell wall	Yes	12
Bsu4037	<i>yycG</i>	1.3. Sensors (signal transduction)		1 (RO-FF-1)
Bsu1596	<i>ftsY</i>	1.6. Protein secretion		2 (RO-E-2)
Bsu3563	<i>yvyH</i>	2.1.1. Specific pathways		3 (RO-E-2, W23, Natto)
Bsu1460	<i>pdhA</i>	2.1.2. Main glycolytic pathways		1 (N10)
Bsu1790	<i>tkt</i>	2.1.2. Main glycolytic pathways		2 (RO-E-2, W23)
Bsu0178	<i>glmS</i>	2.2. Metabolism of amino acids and related molecules		1 (HLR11)
Bsu3073	<i>menC</i>	2.5. Metabolism of coenzymes and prosthetic groups		1 (TU-B-10)
Bsu0606	<i>ydiO</i>	3.2. DNA restriction/modification and repair	Yes	13
Bsu0607	<i>ydiP</i>	3.2. DNA restriction/modification and repair	Yes	13
Bsu0115	<i>rpsJ</i>	3.7.1. Protein synthesis—ribosomal proteins		1 (DV1-F-3)
Bsu4086	<i>rpsR</i>	3.7.1. Protein synthesis—ribosomal proteins		3 (RO-E-2, DV1-E-2, W23)
Bsu3158	<i>mrpD</i>	4.2. Detoxification		1 (DV1-E-2)
Bsu1807	<i>yneS</i>	5.2. From other organisms		1 (TU-B-10)

of sporulation are not completely conserved among all members of *B. subtilis*.

Of the 3,722 genes listed in Table 2, 268 belong to a group of genes that have been designated "essential" in Bsu168 (21). Not surprisingly, most of these genes did not exhibit divergence. Twenty-one of them did, however, reproducibly exhibit a log<sub>2</sub> ratio greater than 1, suggesting that they are divergent among some strains (Table 4). Based on previous reports and an understanding of how some of these "essential" genes function, we can provide an explanation for their loss or divergence among the strains examined. For example, *ydiOP* are essential only when a strain also harbors *ydiR*, *ydiS*, and *ydjA* (26), which together encode the BsuM-specific endonuclease. All of these genes are harbored in a prophage. Any strain that lacks the prophage does not have *ydiR*, *ydiS*, and *ydjA*, and thus *ydiOP* is not required. Also, previous work has shown that the *tag* operon, encoding the enzymes involved in the biosynthesis of one form of teichoic acid (TA), is absent in some members of *B. subtilis* subsp. *spizizenii* (41). In place of *tag* these strains possess a different set of genes that encode a different form of TA (23). Interestingly, one of the proposed methods for distinguishing *B. subtilis* subsp. *subtilis* and *B. subtilis* subsp. *spizizenii* is to characterize the type of TA produced by the strain (25). This approach, however, may not delineate subspecies members as a number of *B. subtilis* subsp. *subtilis* strains also appear to be divergent at this site and one *B. subtilis* subsp. *spizizenii* strain appears to possess all of the genes in the *tag* operon of Bsu168.

Perhaps not surprisingly, more than one-third of "divergent" loci identified in this study have also been predicted to be horizontally transferred (16). Many of these genes, which have G+C content, amino acid, and/or codon bias that deviates from that of the majority of the genome, encode proteins having unknown functions; this group of genes accounts for

nearly 40% of the total number of variable loci. It is worth noting, however, that among all predicted horizontally transferred genes, 145 (30%) never exhibit variability among the strains analyzed. This suggests that these genes are part of the "core" genome (i.e., genes present in all *B. subtilis* strains).

We used Bsu168-specific microarrays to assess the genomic diversity among a collection of *B. subtilis* strains. Our results indicate that as much as 28% of Bsu168's gene content may be missing and/or divergent in this collection of strains. This number is likely to be an underestimate considering that the remaining ~9% of Bsu168's genes not represented on the array are those in suspected prophage regions which would be expected to have an increased incidence of variation. Although it is unclear how much of the variation that we observed using this method represents gene absence rather than gene divergence, we predict that as much as 34% of Bsu168's gene content could be strain-specific or "accessory" genes (i.e., genes not always found in *B. subtilis* isolates). Interestingly, a similar M-CGH study in which 22 *E. coli* isolates were examined using *E. coli* MG1655-specific microarrays revealed that collectively these strains displayed a similar percentage of gene variability (~35%) (15). Whole-genome sequence comparisons of *E. coli* isolates corroborated the M-CGH results; isolates of this species exhibit extensive genomic mosaicism, typified by insertions and deletions of presumably horizontally transferred genes throughout a relatively well-conserved genetic backbone (18, 39). This degree of genomic plasticity has helped explain how *E. coli*, as a species, is able to exploit such a wide variety of niches within the human host. The variability that we have observed in *B. subtilis* genes involved in processes such as sporulation, germination, cell wall synthesis, and antibiotic production may enhance this organism's adaptation to diverse natural environments. It will be interesting to determine how much of this observed variability is driven by horizontal gene

transfer rather than selection for change at the nucleotide level. A recently initiated effort to sequence the genomes of additional *B. subtilis* strains will undoubtedly provide enormous insight into this question, and the results, coupled to the M-CGH data, should increase our understanding of the role that genome diversity plays in the ecology of this nonpathogenic, ubiquitous soil organism.

**Accession numbers.** The nucleotide data for partial *gyrA* sequences were deposited in the NCBI GenBank under accession numbers EF134411 to EF134426. The microarray data for each experiment were deposited in the NCBI GEO database under the Series record GSE 6498.

We thank members of the Kolter lab for many valuable discussions. We also acknowledge Katherine Lemon, Hera Vlamakis, and Vanja Klepac-Ceraj for critical reading of the manuscript.

This work was supported by grants from the NIH (grant GM58213), the Ellison Medical Foundation (grant ID-SS-0248-02), and the DOE (grant DE-FG02-02ER63445) to R.K. A.M.E. was the recipient of a postdoctoral fellowship from the NIH (grant GM072393).

#### REFERENCES

- Auchtung, J. M., C. A. Lee, R. E. Monson, A. P. Lehman, and A. D. Grossman. 2005. Regulation of a *Bacillus subtilis* mobile genetic element by intercellular signaling and the global DNA damage response. *Proc. Natl. Acad. Sci. USA* **102**:12554–12559.
- Binnewies, T. T., Y. Motro, P. F. Hallin, O. Lund, D. Dunn, T. La, D. J. Hampson, M. Bellgard, T. M. Wassenaar, and D. W. Ussery. 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct. Integr. Genomics* **6**:165–185.
- Branda, S. S., F. Chu, D. B. Kearns, R. Losick, and R. Kolter. 2006. A major protein component of the *Bacillus subtilis* biofilm matrix. *Mol. Microbiol.* **59**:1229–1238.
- Branda, S. S., J. E. Gonzalez-Pastor, S. Ben-Yehuda, R. Losick, and R. Kolter. 2001. Fruiting body formation by *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* **98**:11621–11626.
- Branda, S. S., J. E. Gonzalez-Pastor, E. Dervyn, S. D. Ehrlich, R. Losick, and R. Kolter. 2004. Genes involved in formation of structured multicellular communities by *Bacillus subtilis*. *J. Bacteriol.* **186**:3970–3979.
- Burkholder, P. R., and N. H. Giles. 1947. Induced biochemical mutations in *Bacillus subtilis*. *Am. J. Bot.* **34**:345–348.
- Christie, P. J., R. Z. Korman, S. A. Zahler, J. C. Adsit, and G. M. Dunny. 1987. Two conjugation systems associated with *Streptococcus faecalis* plasmid pCF10: identification of a conjugative transposon that transfers between *S. faecalis* and *Bacillus subtilis*. *J. Bacteriol.* **169**:2529–2536.
- Chu, F., D. B. Kearns, S. S. Branda, R. Kolter, and R. Losick. 2006. Targets of the master regulator of biofilm formation in *Bacillus subtilis*. *Mol. Microbiol.* **59**:1216–1228.
- Dorrell, N., S. J. Hinchliffe, and B. W. Wren. 2005. Comparative phylogenomics of pathogenic bacteria by microarray analysis. *Curr. Opin. Microbiol.* **8**:620–626.
- Dziejman, M., E. Balon, D. Boyd, C. M. Fraser, J. F. Heidelberg, and J. J. Mekalanos. 2002. Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proc. Natl. Acad. Sci. USA* **99**:1556–1561.
- Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**:14863–14868.
- El-Helou, E. R. 2001. Identification and molecular characterization of a novel *Bacillus* strain capable of degrading Tween-80. *FEMS Microbiol. Lett.* **196**:119–122.
- Fleischmann, R. D., D. Alland, J. A. Eisen, L. Carpenter, O. White, J. Peterson, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, E. Hickey, J. F. Kolonay, W. C. Nelson, L. A. Umayam, M. Ermolaeva, S. L. Salzberg, A. Delcher, T. Utterback, J. Weidman, H. Khouri, J. Gill, A. Mikula, W. Bishai, W. R. Jacobs, Jr., J. C. Venter, and C. M. Fraser. 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* **184**:5479–5490.
- Fujita, M., J. E. Gonzalez-Pastor, and R. Losick. 2005. High- and low-threshold genes in the Spo0A regulon of *Bacillus subtilis*. *J. Bacteriol.* **187**:1357–1368.
- Fukuya, S., H. Mizoguchi, T. Tobe, and H. Mori. 2004. Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. *J. Bacteriol.* **186**:3911–3921.
- Garcia-Vallve, S., E. Guzman, M. A. Montero, and A. Romeu. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.* **31**:187–189.
- Hemphill, H. E., and H. R. Whiteley. 1975. Bacteriophages of *Bacillus subtilis*. *Bacteriol. Rev.* **39**:257–315.
- Hochhut, B., C. Wilde, G. Balling, B. Middendorf, U. Dobrindt, E. Brzuszkiewicz, G. Gottschalk, E. Carniel, and J. Hacker. 2006. Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic *Escherichia coli* strain 536. *Mol. Microbiol.* **61**:584–595.
- Istock, C. A., K. E. Duncan, N. Ferguson, and X. Zhou. 1992. Sexuality in a natural population of bacteria—*Bacillus subtilis* challenges the clonal paradigm. *Mol. Ecol.* **1**:95–103.
- Joyce, E. A., K. Chan, N. R. Salama, and S. Falkow. 2002. Redefining bacterial populations: a post-genomic reformation. *Nat. Rev. Genet.* **3**:462–473.
- Kobayashi, K., S. D. Ehrlich, A. Albertini, G. Amati, K. K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, S. Aymerich, P. Bessieres, F. Boland, S. C. Brignell, S. Bron, K. Bunai, J. Chapuis, L. C. Christiansen, A. Danchin, M. Debarbouille, E. Dervyn, E. Deuring, K. Devine, S. K. Devine, O. Dreesen, J. Errington, S. Fillinger, S. J. Foster, Y. Fujita, A. Galizzi, R. Gardan, C. Eschevins, T. Fukushima, K. Haga, C. R. Harwood, M. Hecker, D. Hosoya, M. F. Hullo, H. Kakeshita, D. Karamata, Y. Kasahara, F. Kawamura, K. Koga, P. Koski, R. Kuwana, D. Imamura, M. Ishimaru, S. Ishikawa, I. Ishio, D. Le Coq, A. Masson, C. Mauel, R. Meima, R. P. Mellado, A. Moir, S. Moriya, E. Nagakawa, H. Nanamiya, S. Nakai, P. Nygaard, M. Ogura, T. Ohanan, M. O'Reilly, M. O'Rourke, Z. Pragai, H. M. Pooley, G. Rapoport, J. P. Rawlins, L. A. Rivas, C. Rivolta, A. Sadaie, Y. Sadaie, M. Sarvas, T. Sato, H. H. Saxild, E. Scanlan, W. Schumann, J. F. Seegers, J. Sekiguchi, A. Sekowska, S. J. Seror, M. Simon, P. Stragier, R. Studer, H. Takamatsu, T. Tanaka, M. Takeuchi, H. B. Thomaidis, V. Vagner, J. M. van Dijk, K. Watabe, A. Wipat, H. Yamamoto, M. Yamamoto, Y. Yamamoto, K. Yamane, K. Yata, K. Yoshida, H. Yoshikawa, U. Zuber, and N. Ogasawara. 2003. Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* **100**:4678–4683.
- Kozlovskii, I. E., and A. A. Prozorov. 1983. Restriction-modification systems in *Bacillus* strains related to *Bacillus subtilis*. *Genetika* **19**:33–38. [In Russian.]
- Lazarevic, V., F. X. Abellan, S. B. Moller, D. Karamata, and C. Mauel. 2002. Comparison of ribitol and glycerol teichoic acid genes in *Bacillus subtilis* W23 and 168: identical function, similar divergent organization, but different regulation. *Microbiology* **148**:815–824.
- Lindsay, J. A., C. E. Moore, N. P. Day, S. J. Peacock, A. A. Witney, R. A. Stabler, S. E. Husain, P. D. Butcher, and J. Hinds. 2006. Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. *J. Bacteriol.* **188**:669–676.
- Nakamura, L. K., M. S. Roberts, and F. M. Cohan. 1999. Relationship of *Bacillus subtilis* clades associated with strains 168 and W23: a proposal for *Bacillus subtilis* subsp. *subtilis* subsp. nov. and *Bacillus subtilis* subsp. *spizizenii* subsp. nov. *Int. J. Syst. Bacteriol.* **49**:1211–1215.
- Ohshima, H., S. Matsuoka, K. Asai, and Y. Sadaie. 2002. Molecular organization of intrinsic restriction and modification genes BsuM of *Bacillus subtilis* Marburg. *J. Bacteriol.* **184**:381–389.
- Roberts, M. S., and F. M. Cohan. 1995. Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. *Evolution* **49**:1081–1094.
- Roberts, M. S., L. K. Nakamura, and F. M. Cohan. 1996. *Bacillus vallismortis* sp. nov., a close relative of *Bacillus subtilis*, isolated from soil in Death Valley, California. *Int. J. Syst. Bacteriol.* **46**:470–475.
- Salama, N., K. Guillemin, T. K. McDaniel, G. Sherlock, L. Tompkins, and S. Falkow. 2000. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci. USA* **97**:14668–14673.
- Senenschein, A. L., J. A. Hoch, and R. Losick (ed.). 2002. *Bacillus subtilis* and its closest relatives: from genes to cells. ASM Press, Washington, DC.
- Spencer, D. H., A. Kas, E. E. Smith, C. K. Raymond, E. H. Sims, M. Hastings, J. L. Burns, R. Kaul, and M. V. Olson. 2003. Whole-genome sequence variation among multiple isolates of *Pseudomonas aeruginosa*. *J. Bacteriol.* **185**:1316–1325.
- Stover, C. K., X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrenner, M. J. Hickey, F. S. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrook-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. Hancock, S. Lory, and M. V. Olson. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**:959–964.
- Stragier, P., and R. Losick. 1996. Molecular genetics of sporulation in *Bacillus subtilis*. *Annu. Rev. Genet.* **30**:297–341.
- Swofford, D. L. 1998. PAUP: phylogenetic analysis using parsimony. Sinauer Associates, Sunderland, MA.
- Tanaka, T., and T. Koshikawa. 1977. Isolation and characterization of four types of plasmids from *Bacillus subtilis* (natto). *J. Bacteriol.* **131**:699–701.
- Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL\_X windows interface: flexible strategies for

- multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**:4876–4882.
37. **Thorne, C. B.** 1962. Transduction in *Bacillus subtilis*. *J. Bacteriol.* **83**:106–111.
38. **Wang, H., A. P. Roberts, D. Lyras, J. I. Rood, M. Wilks, and P. Mullany.** 2000. Characterization of the ends and target sites of the novel conjugative transposon Tn5397 from *Clostridium difficile*: excision and circularization is mediated by the large resolvase, TndX. *J. Bacteriol.* **182**:3775–3783.
39. **Welch, R. A., V. Burland, G. Plunkett III, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. Mobley, M. S. Donnenberg, and F. R. Blattner.** 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **99**:17020–17024.
40. **Wolfgang, M. C., B. R. Kulasekara, X. Liang, D. Boyd, K. Wu, Q. Yang, C. G. Miyada, and S. Lory.** 2003. Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. USA* **100**:8484–8489.
41. **Young, M., C. Mauel, P. Margot, and D. Karamata.** 1989. Pseudo-allelic relationship between non-homologous genes concerned with biosynthesis of polyglycerol phosphate and polyribitol phosphate teichoic acids in *Bacillus subtilis* strains 168 and W23. *Mol. Microbiol.* **3**:1805–1812.
42. **Zhang, C., M. Zhang, J. Ju, J. Nietfeldt, J. Wise, P. M. Terry, M. Olson, S. D. Kachman, M. Wiedmann, M. Samadpour, and A. K. Benson.** 2003. Genome diversification in phylogenetic lineages I and II of *Listeria monocytogenes*: identification of segments unique to lineage II populations. *J. Bacteriol.* **185**:5573–5584.