Small Genes under Sporulation Control in the *Bacillus subtilis* genome \(^*\)†

Matthias Schmalisch,1 Elisa Maiques,1 Lachezar Nikolov,1 Amy H. Camp,1 Bastien Chevreux,2 Andrea Muffler,2 Sabrina Rodriguez,3 John Perkins,3 and Richard Losick1*

The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138; DSM Nutritional Products Ltd., Kaiseraugst, Switzerland; and DBC, DFS Food Specialties B.V., Delft, Netherlands

Received 10 May 2010/Accepted 28 July 2010

Using an oligonucleotide microarray, we searched for previously unrecognized transcription units in intergenic regions in the genome of *Bacillus subtilis*, with an emphasis on identifying small genes activated during spore formation. Nineteen transcription units were identified, 11 of which were shown to depend on one or more sporulation-regulatory proteins for their expression. A high proportion of the transcription units contained small, functional open reading frames (ORFs). One such newly identified ORF is a member of a family of six structurally similar genes that are transcribed under the control of sporulation transcription factor α\(^*\) or σ\(^*\). A multiple mutant lacking all six genes was found to sporulate with slightly higher efficiency than the wild type, suggesting that under standard laboratory conditions the expression of these genes imposes a small cost on the production of heat-resistant spores. Finally, three of the transcription units specified small, noncoding RNAs; one of these was under the control of the sporulation transcription factor σ\(^*\), and another was under the control of the motility sigma factor σ\(^{II}\).

*Bacillus subtilis*, also known as the hay bacillus and commonly found in soil, is the best-characterized member of the Gram-positive group of bacteria. It is highly accessible for genetic manipulation and has therefore served as a model organism for laboratory studies, especially for cell differentiation processes (for reviews see references 1 and 29). With the publication of its genome sequence by Kunst et al. in 1997 (18), it also became the first Gram-positive bacterium with a complete genome sequence. At the time of publication, 4,100 protein-coding genes were annotated in a total of 4,214,810 bp. These protein-coding genes constitute 87% of the genome, a number which has been found to be very similar to the composition of other bacterial genomes such as the genome of *Escherichia coli* (4). Although the ratio of noncoding DNA to coding DNA is very consistent in bacterial genomes, ranging from 6 to 14% of noncoding DNA for the majority of sequenced genomes (28), it has become apparent in recent years that the percentage of coding RNAs is actually higher than previously anticipated.

In the last decade a wealth of additional genes has been discovered in previously annotated genomes that play important roles in a range of cellular processes. Most of these genes encode small regulatory RNAs (sRNAs or ncRNAs, for noncoding RNAs; one of these was under the control of the sporulation transcription factor α\(^*\) and translation (for a microreview on OxyS and its target *rpoS*, see reference 26). Small RNAs for a number of other bacteria have been discovered, including both Gram-positive as well as Gram-negative species. Examples are *Staphylococcus aureus* (16), *Pseudomonas aeruginosa* (45) and *Clostridium acetobutylicum* (5). In addition to experimentally verified small RNAs, several hundred RNAs are predicted by computational methods. Using a novel prediction algorithm, Tran et al. (41) reported the finding of 601 potential small RNAs in *E. coli*.

In addition to small RNAs, the focus of recent microbial genome studies has also been on small ORFs in intergenic regions. The identification or discovery of small ORFs (for example, those of less than 50 codons) can be challenging. They often cannot be reliably annotated, and their products are difficult to detect with standard biochemical methods, such as two-dimensional gel electrophoresis. Applying new computational models for sequence comparison and ribosome binding to the genome of *E. coli*, Hemm et al. (15) discovered 18 additional small ORFs and confirmed the synthesis of 20 previously annotated proteins smaller than 50 amino acids. Previously unannotated ORFs have also been reported in *Pseudomonas fluorescens*, where 16 new ORFs where found, with 9 of them being smaller than 100 amino acids (17).

In *B. subtilis*, several new small RNAs have been discovered in recent years, including RatA, an antisense RNA for a peptide toxin (35), FsrA, a small RNA regulated by the iron-response regulator Fur (12), SR1, a small RNA controlled by the gluconeogenesis regulator CcpN (20), and several novel small RNAs under sporulation control (36). In addition to these noncoding RNAs, additional small
ORFs in previously “empty” intergenic regions have been discovered. Perhaps the first to be identified and still one of the most striking examples is the 2.6-codon-long spoVIM gene, which plays an essential role in spore formation (8). Gaballa et al. (12) characterized three small ORFs, of which two were previously not annotated: fbpB (intergenic region between fbpA and ydbM) and fbpC (between ypbQ and ypbR). Another example is mciZ, a small 40-codon ORF that was discovered by Handler et al. (14) in the intergenic region of nudF and ykpL and that has been shown to interact with the cell division protein FtsZ. In addition to these findings, genome-wide transcriptional studies have led to the discovery of previously unknown genes. Lee et al. (19) used an antisense microarray that covered not only known ORFs but also some intergenic regions. RNA expression analysis suggested 35 independent transcripts in intergenic regions, with 20 of them containing putative ORFs. Finally, the most recent findings come from a genome-wide transcriptional study done by Rasmussen et al. (25). This group used a microarray to investigate all transcriptionally active regions in\emph{ B. subtilis}. Observed transcriptional activity in intergenic regions suggests 84 putative small RNAs. To date, none of these has been confirmed by other methods.

The focus of this investigation was a transcriptional analysis of the intergenic regions of \emph{B. subtilis}, with a special emphasis on unknown small genes that are turned on during sporulation. By carrying out analysis on a genome-wide basis, we sought to extend earlier work in which several genes for small RNAs that are transcribed in response to sporulation (36) were discovered. For this purpose, we used a densely tiled microarray that covered all intergenic regions larger than 50 nucleotides. As we report, 19 transcript units were identified, 11 of which proved to be novel. For this purpose, we used a densely tiled microarray that covered all intergenic regions larger than 50 nucleotides. As we report, 19 transcript units were identified, 11 of which proved to be novel. For this purpose, we used a densely tiled microarray that covered all intergenic regions larger than 50 nucleotides. As we report, 19 transcript units were identified, 11 of which proved to be novel.
600 nm. Bioluminescence is reported in arbitrary units. Of note, we found that cell growth/sporulation on agar pads, as opposed to in liquid medium, resulted in more robust and reproducible luciferase measurements.

Heat kill assay. To quantify spore formation, cells were induced to sporulate by nutrient exhaustion in Difco sporulation medium (DSM) (30). A single colony was picked from a fresh agar plate and used to inoculate 5 ml of DSM and allowed to grow at 37°C for 24 h. The frequency of sporulation was calculated from the number of CFU before and after heat treatment (20 min at 80°C), which was determined by plating various dilutions on agar plates. For each experiment, wild-type B. subtilis PY79 was used as a control, and for comparison the number of spores obtained with the wild type (1 × 10⁸ to 5 × 10⁷ per ml) was set as 100%.

Fluorescence microscopy. For the localization of the sporulation-dependent GFP fusion proteins, cultures of the different strains were grown overnight in LB medium and diluted in medium used for resuspension. At mid-exponential phase, sporulation was induced by resuspension in SM medium (38). Cells were harvested at different times after induction of sporulation by centrifugation and resuspended in phosphate-buffered saline (PBS) buffer. Where necessary, membranes were visualized by adding the membrane dye FM4-64 [N-[3-sulfoethyl]1,3-bis(3-[(diethylamino)methyl]carbonyl]-1H-pyridinium dibromide; green fluorescent] [Invitrogen] or TMA-DPH (1-[4-fluoromethyl]-1H-pyridinium 6-phenyl-1,3,5-hexatriene; blue fluorescent] (Molecular Probes) to the PBS buffer (final concentration, 1 μg/ml). For the fluorescence microscopy, cells were immobilized on a 1% agarose pad covered with a polylysine-treated coverslip. The same equipment and methods were used as described in previous work by this lab (see reference 11 for details).

β-Galactosidase assay. For measurements of gene activity with lacZ reporter genes, cells were grown in LB medium or induced to sporulate by the resuspension method (23, 38). Aliquots of cells were collected at intervals, stored at −20°C, and processed for β-galactosidase activity essentially as previously described (6).

Competition experiments. To identify a sporulation phenotype of mutant strains lacking one or more proteins of the yrhJ-yrhK homologs, including the newly identified gene in the yrhJ-yrhK intergenic regions, we challenged these strains against the parent strain in a competition experiment. To identify the location of the strains on an agar plate, an isopropyl-β-D-thiogalactopyranoside (IPTG)-inducible β-galactosidase gene was utilized. For each experiment, this gene was present in either the wild-type or the mutant strain. At the beginning of the experiment, optical densities (OD) of cell cultures of the wild type and of the mutant strain grown overnight in LB medium were measured. Equal amounts (1 OD unit) of both cell cultures were mixed and diluted 1:1,000 in 10 ml of DSM. These steps were repeated at least five times, and dilutions of the cell cultures were plated on LB agar plates containing 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal; 40 μg/ml). Numbers of blue and white colonies from the number of CFU before and after heat treatment (20 min at 80°C), which was determined by plating various dilutions on agar plates. For each experiment, wild-type B. subtilis PY79 was used as a control, and for comparison the number of spores obtained with the wild type (1 × 10⁸ to 5 × 10⁷ per ml) was set as 100%.

RNA isolation. Cell cultures for RNA isolation were harvested by a brief centrifugation step (5 min at 8,500 rpm at room temperature), and the pellet was immediately frozen in liquid nitrogen and kept at −80°C until further processing. The pellet was resuspended in 1 ml of lysis buffer containing 500 μl of phenol-chloroform and 100 μl of 10% sodium dodecyl sulfate. The solution was transferred into lysis tubes (MP Biomedicals) containing glass beads (Qiobio) and processed in a bead beater (Qiobio) for 10 s at 6,600 rpm. The lysis tubes were centrifuged at 13,000 rpm at 4°C, and supernatant without cell debris or the matrix was transferred to the microcentrifuge tubes and phenol-chloroform steps with 1 ml RNA was precipitated by addition of an equal volume of ethanol and incubation at −20°C overnight. RNA was pelleted by centrifugation (30 min at 13,000 rpm at 4°C) and washed twice with 1 ml of 70% ethanol. The pellet was dried and resuspended in 100 μl of diethyl pyrocarbonate (DEPC)-water. RNA was treated with DNAse I to remove residual chromosomal DNA and purified with either a Qiagen RNeasy minikit or Ambion mirVana microRNA (miRNA) for small-RNA enrichment) according to the manufacturer’s recommendations. Quality of the RNA preparation was verified by using an Agilent Bioanalyzer.

cDNA labeling, hybridization, and microarray scanning. Preparation of cDNA targets and hybridization were based on methods described previously (19). Washing and staining with GeneChip were performed according to Affymetrix’s standard protocol. Microarrays were processed with an Affymetrix scanner and analyzed by using the Affymetrix gene expression analysis suite. Data were normalized according to the mean of the sum of all of the comparable experiments (robust multichip average [RMA] method). Expression data were visualized using the Affymetrix Integrated Genome Browser.

Homolog search and alignment. The ORFs in the intergenic regions were queried by using the protein sequence against the translated nucleotide database (BLASTN [http://www.ncbi.nlm.nih.gov/BLAST]). The homologs for the ORF in the yrhJ-yrhK intergenic regions were found by using protein-protein BLAST against the B. subtilis genome only. The protein alignment for yrhJ and yrhK were used and the multiple alignments for the yrhJ homologs were done using the ClustalW2 algorithm and the standard parameters (http://www.ebi.ac.uk/Tools/clustalw2/index.html).

RESULTS AND DISCUSSION

Identifying transcripts in intergenic regions. We developed an oligonucleotide microarray using Affymetrix GeneChip technology to search for transcripts in intergenic regions of the B. subtilis genome. The microarray was designed to include all intergenic regions with a length of at least 50 nucleotides between the stop and the start codons of the neighboring genes. Regions containing tRNA or rRNA genes were excluded from the microarray. Hybridization experiments were carried out with RNA isolated from the prototrophic wild-type strain PY79. Cells were grown in sporulation medium and harvested at the mid-exponential phase of growth at 2 or 5 h after the onset of sporulation. We enriched for small RNAs, as described in
Materials and Methods, and then created fluorescently labeled cDNAs by reverse transcription. Finally, the fluorescent probes were annealed with the microarray (for details, see Materials and Methods). Analysis of the microarray data showed hybridization signals above background in most intergenic regions (>75%). We used three criteria in an initial effort to distinguish transcripts that arose from initiation within an intergenic region from the 3' or 5' untranslated regions of adjacent genes: (i) the minimum length of the transcript was 30 bp as judged by having five contiguous hybridization signals above background, (ii) the distance of the transcript from flanking genes was greater than 30 bp if the adjacent gene was transcribed from the same DNA strand, and (iii) no known or predicted transcription termination sequence was present in the intergenic region.

Thirty-four transcripts that met these criteria and that in addition appeared to be present after the onset of sporulation but absent in control experiments with RNA from cells isolated late in stationary phase in LB medium (data not shown) were chosen for further study. As a next step, we asked whether the newly identified transcripts were in these intergenic regions: (i) absent in control experiments with RNA from cells isolated late in stationary phase in LB medium (data not shown) were (ii) the distance of the transcript from flanking genes was greater than 30 bp if the adjacent gene was transcribed from the same DNA strand, and (iii) no known or predicted transcription termination sequence was present in the intergenic region.

Table 1 lists the intergenic regions for which we identified an apparent promoter.

Recently, Rasmussen et al. (25) globally identified regions of the B. subtilis genome that are transcriptionally active during growth in minimal medium or LB medium. Our experiments were designed to detect genes in intergenic regions that are induced in DS sporulation medium. Nonetheless, 8 out of the 19 intergenic transcription units identified in our investigation were also detected in the study of Rasmussen et al. The newly identified transcripts were in these intergenic regions:

**Small ORFs**. We looked for possible ORFs in the intergenic regions containing transcripts. If a putative ORF was found, we investigated whether the predicted amino acid sequence was homologous to amino acid sequences in the databases using tBLASTN (NCBI). ORFs were found in 16 of the intergenic regions.

### Table 1. Novel genes and homologs

<table>
<thead>
<tr>
<th>Gene orientation</th>
<th>Left gene</th>
<th>Right gene</th>
<th>Size</th>
<th>Homolog(s)</th>
<th>Previous report(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>← ← ←</td>
<td>yhbC</td>
<td>yhbB</td>
<td>28 aa</td>
<td>B. licheniformis</td>
<td></td>
</tr>
<tr>
<td>← ← ←</td>
<td>yhbS</td>
<td>cotF</td>
<td>49 aa</td>
<td>B. licheniformis, B. amylooliquefaciens</td>
<td>3, 25</td>
</tr>
<tr>
<td>← ← ←</td>
<td>yrpD</td>
<td>yrpE</td>
<td>25 aa</td>
<td>G. thermodenitrificans, B. amylooliquefaciens, B. pumilis</td>
<td>3, 25</td>
</tr>
<tr>
<td>← ← ←</td>
<td>yrkD</td>
<td>yrkC</td>
<td>79 aa</td>
<td>G. kaustophilus, G. uraniumreducens, D. psychrophila</td>
<td>3, 25</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>yozJ</td>
<td>rapK</td>
<td>90 aa, 56 aa</td>
<td>B. amylooliquefaciens</td>
<td>3, 25</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>yrpJ</td>
<td>ytpJ</td>
<td>63 aa</td>
<td>B. pumilis, B. licheniformis, B. halodurans, O. iheyensis, E. sibiricum, H. modesticaldum</td>
<td>3, 25</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>purD</td>
<td>yezC</td>
<td>44 aa</td>
<td>B. anthracis, B. cereus, B. amylooliquefaciens, L. sphaericus</td>
<td>3, 25</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>yuwC</td>
<td>yuwB</td>
<td>55 aa</td>
<td>B. licheniformis, B. pumilis</td>
<td>3, 25</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>ypeP</td>
<td>ypeS</td>
<td>48 aa</td>
<td>B. amylooliquefaciens, B. licheniformis, G. kaustophilus</td>
<td>3, 25</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>yoeL</td>
<td>yoeM</td>
<td>75 aa</td>
<td>B. subtilis, B. amylooliquefaciens, B. pumilis</td>
<td>3</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>ydcJ</td>
<td>ydcK</td>
<td>37 aa</td>
<td>B. licheniformis, B. amylooliquefaciens, B. anthracis, B. coagulans</td>
<td>3</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>ydzH</td>
<td>ydfR</td>
<td>76 aa, 101 aa</td>
<td>B. subtilis, B. amylooliquefaciens, B. licheniformis, G. kaustophilus, G. thermodenitrificans, B. pumilis</td>
<td>3</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>ynzH</td>
<td>thyA</td>
<td>58 aa</td>
<td>B. halodurans, G. kaustophilus, B. clausii, B. cereus</td>
<td>3, 25</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>yoaH</td>
<td>yoaI</td>
<td>39 aa</td>
<td>B. halodurans, G. kaustophilus, B. clausii, B. cereus</td>
<td>3</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>yjgH</td>
<td>yjgH</td>
<td>55 aa</td>
<td>B. amylooliquefaciens, B. licheniformis, G. kaustophilus, B. pumilus</td>
<td>3</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>yrhK</td>
<td>yrhJ</td>
<td>100 nt</td>
<td>B. amylooliquefaciens</td>
<td>3</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>ykaJ</td>
<td>ykaJ</td>
<td>120 nt</td>
<td>B. pumilis, B. licheniformis</td>
<td>12</td>
</tr>
<tr>
<td>← ← ← ←</td>
<td>yocG</td>
<td>yocH</td>
<td>220 nt</td>
<td>B. pumilis, B. licheniformis</td>
<td>12</td>
</tr>
</tbody>
</table>

a Outside arrows represent left and right known genes. Internal arrows indicate newly identified genes based on luciferase or GFP fusion studies; in two cases two new genes were identified.
b Size of proteins predicted from ORFs in the IGR. In two cases more than one ORF was detected. In the last three examples no coding region was identified, but the size of the putative transcript is provided. aa, amino acid; nt, nucleotide.
c Similarity determined at the DNA sequence level. In two cases paralogs to B. subtilis genes (bold) were detected. In two other cases, no similarities were found (—).

The following are Bacillus species: B. licheniformis, B. amylooliquefaciens, B. pumilis, B. halodurans, B. anthracis, B. cereus, B. clausii, B. coagulans, B. weihenstephanensis, B. thuringiensis, B. megaterium, B. licheniformis. Other organisms are as follows: D. psychrophila, Desulfoaltera psychrophila; E. sibiricum, Geobacillus safensis, G. kaustophilus, Geobacillus subtilis, G. thermodenitrificans, Geobacillus thermodenitrificans; G. uraniumreducens, Geobacter uraniumreducens; H. modesticaldum, Helobacterium modesticaldum; L. sphaericus, Lysinibacillus sphaericus; O. iheyensis, Oceanobacillus iheyensis.

The bacterial luciferase operon (luxABCDE) that was modified to harbor optimal translational signals for expression in B. subtilis (24; also the present study). Of the 29 fusions, 10 failed to show promoter activity under all conditions tested (for a list of the fusions with no activity, see Table S3 in the supplemental material). In contrast, the remaining 19 fusions exhibited luciferase activity in either sporulation medium (DS) and/or in complex medium (LB), indicating the apparent presence of promoters in these intergenic regions (see Fig. S1, gray solid lines, in the supplemental material). Table 1 lists the intergenic regions for which we identified an apparent promoter.
regions, with 15 of the 16 ORFs having sequence similarity to ORFs in other species (Table 1). Two of these intergenic regions, yozJ-rapK and ydzH-ydfR, each contained two putative ORFs. While both ORFs in the ydzH-ydfR intergenic region showed high similarity to known or predicted proteins in other species, only one of the two putative ORFs in the yozJ-rapK region has a known homolog. Thus, we identified a total of 18 putative protein-coding sequences. Figure 1 shows the beginning and end of the ORFs in their genetic context. For the two cases of intergenic regions containing tandem pairs of ORFs (yozJ-rapK and ydzH-ydfR), no promoter activity was detected for the region between the first and second ORFs (data not shown). Therefore, in both cases the pairs of ORFs appear to constitute single transcriptional units (operons).

During the course of these experiments, Barbe et al. (3) reported a reannotation of the B. subtilis genome sequence, which included ORFs in intergenic regions where none had been previously annotated. As a result, 11 of the 18 putative ORFs identified by us are also present in the most current B. subtilis genome annotation (http://genodb.pasteur.fr/cgi-bin/WebObjects/GenoList). All of these ORFs are either homologs to genes of unknown function or do not have homologs to previously reported genes (3). Because the newly annotated ORFs were not experimentally confirmed, we simply refer to them according to the intergenic region in which they are located. However, to facilitate a cross-comparison, we have included in Fig. 1 the new annotation names for the 11 cases as they appear in the annotation of Barbe et al. (3).

To investigate whether the newly discovered ORFs were indeed functional and translated into proteins, we fused the promoter regions including the first 8 to 10 codons of each ORF (only the first ORF in the case of the two tandem pairs of ORFs) in frame to a variant of the luciferase operon in which the RBS and start codon of the first gene (luxA) were absent. In this way, we created translational fusions for all but three (yxbC-yxbB, yozJ-rapK, and yrkD-yrkC) of the 16 ORFs. Luciferase activity was observed for all 13 of these fusions (see Fig. S1, dotted gray lines, in the supplemental material), results that confirm that these ORFs are indeed translated. However, some of the translational luciferase fusions showed low activity relative to their corresponding transcriptional fusion (ytrI-ytqI, purD-yczG, and ypcP-ypbS) (see Fig. S1), whereas other translation fusions showed activity comparable to their corresponding transcriptional fusion (see Fig. S1). Because the luxA gene in the transcriptional fusions had an optimal RBS for B. subtilis (22), these findings seem to reflect variations in the strength of the translational signals for the ORFs. Putative RBSs and start and stop codons for each ORF are shown in Fig. 1.

In general, we observed weaker translational activity for ORFs beginning with GTG and TTG start codons than for those beginning with ATG. These findings are in agreement with previous results in B. subtilis, showing decreased translation initiation and mRNA stability associated with GTG and TTG start codons (33). An apparent exception was the ynzH-thyA translational fusion, which, despite its GTG start codon, showed almost the same level of activity for both the transcriptional and translational fusions (Fig. S1). However, in this case, the putative RBS (the sequence TTGGAGG located 7 bp upstream of the start codon) (Fig. 1) is close to the optimal for B. subtilis (22).
Finally, and of note, there is one exceptional case. The ORF in the ywqC-ywqB intergenic region exhibited no apparent RBS upstream of the start codon (Fig. 1); however, the translational fusion showed high expression levels (see Fig. S1 in the supplemental material). No other potential start codons or RBS could be identified. Expression of ywqC-ywqB in the absence of an apparent RBS was confirmed with an in-frame fusion to the gene for the green fluorescent protein (Fig. 2; IGR CB-GFP).

Also setting a boundary on the region in which translation could commence was the cloned segment of DNA that was used for the promoter fusion. This DNA segment contained a consensus match for a \( \sigma^E \)-dependent promoter located 18 bp upstream of the start codon, reinforcing the conclusion that the putative ATG start codon is, indeed, the translation start point.

**Sporulation-dependent expression.** As indicated above, the intergenic regions that were chosen for further study specified transcripts that were preferentially produced during sporulation and, hence, candidates for genes under sporulation control. To further investigate whether production of these transcripts was indeed under sporulation control, the transcriptional fusions were tested for luciferase activity in a series of mutants for various sporulation-regulatory proteins. Entry into sporulation is governed by the response regulator SpooA and the alternative sigma factor \( \sigma^H \) (37). SpooA is active in the predivisional sporangium but later, after the stage of asymmetric division, becomes active selectively in the mother cell. Subsequent gene expression is governed by the sequential appearance of the sigma factors \( \sigma^E \), \( \sigma^F \), \( \sigma^G \), and \( \sigma^K \), with \( \sigma^E \) and \( \sigma^G \) being specific to the forespore and \( \sigma^H \) and \( \sigma^K \) being specific to the mother cell (39).

Using the above described reporter strains, we found that 11 of the newly identified promoters were under sporulation control; i.e., expression was at least dependent on SpooA. These 11 included one putative, non-protein-coding gene, yoCG-yocH, and 10 protein-coding genes (Table 2). Transcription from promoters for three of the genes (those contained in the ywqC-ywqB, yjbG-yjbH, and yqeN-comEC regions) was dependent on SpooA but not on any of the sporulation sigma factors. Of the remaining eight, promoter activity for three protein-coding genes (ydeH-ydjR, yocL-yocM, and ytrf-ylql) as well for the noncoding gene (yoCG-yocH) was dependent on both SpooA and \( \sigma^F \). (We consider the yoCG-yocH gene further below.) Finally, promoter activity for two genes, those in the intergenic regions ypcP-ypbS and ynzH-thyA, were dependent on only \( \sigma^G \) and \( \sigma^K \), respectively. For the remaining two cases, those of the promoters for ybS-cofF and yoaH-yoaI, expression depended on SpooA, but additional dependencies on sporulation sigma factors were not investigated.

Reinforcing the results from the dependency experiments were images from fluorescent microscopy experiments with in-frame fusions of newly discovered protein-coding genes to the gene for GFP. For example, Fig. 2 shows that the SpooA-dependent gene in the ywqC-ywqB intergenic region was expressed in the predivisional sporangium. Likewise, and as expected, expression of the \( \sigma^E \)-dependent ytrf-ylql gene was limited to the mother cell. In contrast, expression of the \( \sigma^G \)-dependent gene contained in the ypcP-ypbS region was restricted to the forespore. Finally, and in contrast to the whole compartment fluorescence patterns seen for the above fusion proteins, GFP fused to the \( \sigma^E \)-dependent gene contained in the ynzH-thyA region localized in a ring around the forespore (Fig. 2, the forespore at the far right). This is a localization pattern characteristic of coat proteins. It seems likely, therefore, that the product of the newly discovered ORF in the ynzH-thyA region is a previously unrecognized component of the coat.

Finally, we return to the eight genes whose expression did not exhibit a dependence on SpooA. Thus, despite the results from the microarray data (stationary-phase expression in DS but not in LB medium), these genes appeared not to be under sporulation control. Instead, we conclude that these are genes

![FIG. 2. Cell-specific expression of newly identified genes under sporulation control. Fluorescent micrographs of cells expressing GFP fused in frame to the 3′ terminus of the indicated genes. Sporulation was induced by resuspension in SM medium. Cells were stained with the membrane dye TMA-DPH (Molecular Probes). Shown is a merged image of the GFP (green) and the membrane dye (blue). The ywqC-ywqB gene (IGR CB-GFB), which is under SpooA control, was expressed in the predivisional sporangium (visualized at hour 1); the ytrf-ylql gene (IGR II-GFP), which is under \( \sigma^E \) control, and ynzH-thyA (IGR HA-GFP), which is under \( \sigma^G \) control, were expressed in the mother cell (visualized at hours 3 and 5, respectively); and ypcP-ypbS (IGR PS-GFP), which is under \( \sigma^K \) control, was expressed in the forespore (visualized at hour 3).](http://jb.asm.org/)
that are simply more actively expressed in the postexponential phase of growth in DS medium.

Paralogs to newly identified genes. As shown in Table 1 and discussed earlier, two of the newly identified protein-coding genes have apparent paralogs in the *B. subtilis* genome. The first ORF in the *ydzH-ydfR* intergenic region shares significant sequence similarity with *ypzD*. An alignment of the predicted products of the two genes reveals 47% identical and 67% similar amino acids over the entire length of the two proteins (Fig. 3B). Reinforcing the striking similarity in protein-coding sequences, the two genes are regulated in a similar manner. This was demonstrated by constructing in-frame fusions of *ypzD* to the luciferase operon and *gfp* gene. The results with fusions showed that, as with the *ydzH-ydfR* ORF (above), expression of *ypzD* was dependent on σ^E (data not shown) and expressed specifically in the mother cell (Fig. 3A). It seems likely that the *ydzH-ydfR* and *ypzD* genes play similar roles in sporulation.

Even more striking is the case of the newly identified gene in the *yocL-yocM* intergenic region. Its predicted product has significant similarity to the proteins encoded by the *B. subtilis* genes: *yocN*, *yozN*, *ydzH*, *ydgA*, and *ydgB*. Although all six predicted protein products share a high number of conserved amino acids (Fig. 4B), neighbor-joining shows that the six proteins subdivide into two groups. The predicted product of the newly identified gene in the intergenic region of *yocL-yocM* shows the highest similarity to the *yocN* and *yozN* gene products, with 59% and 49% similar amino acids, respectively, while similarity to the other three inferred gene products was below 40%. These three, *ydzH*, *ydgA*, and *ydgB*, appear more phylogenetically related to each other, with a sequence similarity greater than 60%.

Because *yocL-yocM* was identified as being under sporulation control, we asked if any of the apparent paralogs are also involved in sporulation. Two of the genes, *yocN* and *yozN*, have been previously identified as being under the control of the sporulation-dependent sigma factor σ^E and being expressed in a mother cell-specific fashion (7). To investigate whether the remaining three genes are also involved in sporulation, we constructed in-frame fusions to both the luciferase operon and the *gfp* gene. Because *ydgB* and *ydgA* seem to be in an operon (their ORFs are separated only by 13 bp), we examined the expression of only the upstream member of the operon, *ydgB*. The luciferase operon was fused to the upstream region of *ydzH* and *ydgB*, and *gfp* was fused to the 3’ ends of *yozN*, the *yocL-yocM* IGR, *ydgA*, and *ydzH* genes. Interestingly, luciferase fusions to *ydzH* and *ydgB* were switched on later in sporulation than the other three paralogs. Furthermore, dependence studies showed that *ydgB* and *ydzH* were under the control of the late-appearing, mother cell-specific sigma factor σ^K (data not shown). Finally, fluorescent microscopy experiments with the *gfp* fusions confirmed, as expected, that expression was confined to the mother cell (Fig. 4A). In addition, in the cases of *ydgA* and *ydzH*, striking punctate patterns of localization were observed in which one or more foci were seen around the forespore (Fig. 4A). Such a punctate pattern of localization has
been seen previously for GFP fusions to other proteins produced in the mother cell and suggests that YdgB and YdzH are likely to be components of the spore coat (42).

A multiple mutant exhibits enhanced sporulation. We built deletion mutants for the 11 newly identified genes under sporulation control (the 10 protein-coding genes and the putative noncoding gene) as well as for the four additional paralogous genes (ypzD, ydgA, ydgB, and ydzH) herein found to be under sporulation control. No conspicuous defect in the production of heat-resistant spores (see Table S4 in the supplemental material) or in the timing of sporulation, as judged by microscopy (data not shown), was detected for any of the mutants. Because the newly discovered gene in the yocL-yocM intergenic region has five paralogs, we constructed mutants for all six genes individually and together. Neither mutants of individual genes alone nor a multiple mutant lacking all six genes exhibited a significant defect in sporulation (see Table S5 in the supplemental material).

As a more sensitive test for a subtle defect in growth or sporulation, we carried out competition experiments between the wild type and various mutants through cycles of growth and sporulation. To distinguish the wild type from the mutants, we inserted an IPTG-inducible lacZ gene in either the wild type or the mutants. The experiment was started by mixing exponentially growing wild-type and mutant cells and inoculating Difco sporulation medium. After 24 h, the nonsporulated cells were inactivated (at 80°C for 20 min), and fresh Difco sporulation medium was inoculated (1:1,000). The ratio of the two strains was determined by plating dilutions of the culture on plates with X-Gal and IPTG, with either the wild type or mutant showing up as blue colonies. Experiments with five different mutants (ywqC-ywqB, ydcI-ydcK, yocG-yocH, ypcP-ypbS, and ynzH-thyA) exhibited no differences in growth and sporulation between the wild type and the mutants (data not shown).

However, the multiple mutant lacking all six genes, similar to yocL-yocM, exhibited a phenotype. Figure 5A shows the ratios of wild-type to mutant cells through six cycles of sporulation, with the mutant harboring the lacZ marker. Surprisingly, the multiple mutant out-competed the wild-type cells. The presence of the marker gene had no effect on the experiment: inserting lacZ into the parent and repeating the competition experiment again showed that the multiple mutant out-competed the wild type (data not shown).

The competition advantage was specific for sporulation because growth in nonsporulation medium, LB medium, showed no advantage of the mutant over the wild type (Fig. 5B). To further investigate whether the ability of the multiple mutant to out-compete the wild type occurred during the growth phase of the cycles, we carried out an additional competition experiment, this time keeping the cells continually growing in Difco sporulation medium by continuous transfer while in the exponential phase of growth. Under these continuous growth con-
follows.

The gene for the second noncoding RNA lies between \textit{yrhJ} and \textit{yrhK} in the same orientation as \textit{yocH} and opposite to \textit{yrhJ} (Table 1 and Fig. 6A). No ORF is present between \textit{yrhJ} and \textit{yrhK}, and the gene must therefore specify a noncoding RNA. Strong promoter activity was detected downstream of \textit{yrhJ}. This promoter was active during the exponential phase of growth, shutting down at the beginning of stationary phase (see Fig. S1, \textit{yrhJ-yrhK}, in the supplemental material). In preliminary work, we carried out a rapid amplification of cDNA ends (RACE) experiment to attempt to identify the promoter for the \textit{yrhJ-yrhK} RNA (data not shown). We noticed that the region so identified contained a perfect match to the consensus
control of $\sigma^E$. Likewise, and even more striking, the ORF in the yokE-yocH intergenic region is apparently paralogous to five other genes; two of these were known to be under sporulation control, and the remaining three were here shown (two inferred to be members of the same operon) to be under sporulation control. Thus, our analysis brings to 15 the number of genes newly recognized as being under sporulation control.

A multiple mutant lacking the ORF in the yokL-yocM intergenic region and its five paralogs exhibited a novel phenotype. Instead of being impaired in spore formation, as might have been anticipated, it produced spores more efficiently than the wild-type parent in competition experiments involving cycles of spore formation. We conclude that under laboratory conditions, expression of the six paralogs imposes a small cost to spore formation. Presumably, under natural conditions this cost is offset by a fitness benefit that was not replicated in the laboratory.

A final noteworthy feature of this investigation was the identification of three small, non-protein-coding RNAs, two of which are newly identified. One of the newly identified sRNAs (the yokG-yocH RNA) was under the control of the sporulation sigma factor $\sigma^E$, whereas the other (the yrhJ-yrhK RNA) was under the control of the motility sigma factor $\sigma^D$. A striking feature of the yrhJ-yrhK RNA is its conspicuous complementarity to multiple protein-coding genes, raising the possibility that it influences the stability or translation of the corresponding mRNAs.

ACKNOWLEDGMENTS

We thank Thomas Albert for assistance in the preparation of the Bacillus sRNA microarray and Markus Wyss for helpful comments on the paper.

This work was supported by DSM and by NIH grant GMA18568 to R.L.M. was supported by a postdoctoral fellowship from the DFG (German Research Foundation). A.H.C. was supported by a Helen Hay Whitney postdoctoral fellowship. E.M. is the recipient of an MEC postdoctoral fellowship from the Secretaría General de Estado de Universidades e Investigación del Ministerio de Educación y Ciencia (Spain).

REFERENCES


