The transcriptional profile of early to middle sporulation in *Bacillus subtilis*

Paul Fawcett*†‡§, Patrick Eichenberger†, Richard Losick†¶, and Philip Youngman*

*Millennium Pharmaceuticals, Cambridge MA, 02138; †Harvard University, Department of Molecular and Cellular Biology, Cambridge, MA, 01238; and ‡University of Georgia, Department of Genetics, Athens GA, 30602

Contributed by Richard Losick, May 9, 2000

Spore formation by *Bacillus subtilis* is governed by global changes in gene transcription. We used nylon-substrate DNA arrays representing ≈96% of the predicted open reading frames in the *B. subtilis* chromosome to compare the pattern of transcripts from wild-type cells with the pattern from cells mutant for the sporulation transcription factors Spo0A or σ^F. We found 520 genes whose transcript levels were at least 3-fold dependent on Spo0A but not on σ^F, and an additional 66 genes whose transcript levels were dependent upon both regulatory proteins. Two strategies were used to help assign genes to the direct control of a particular developmental regulatory protein. In one approach, we analyzed the effects on global gene expression of artificially producing a constitutively active form of Spo0A during growth. In a second approach, Hidden Markov models were used to identify promoters likely to be activated by Spo0A, σ^F, or a third sporulation transcription factor, σ^E. In addition to detecting known sporulation genes, we identified many genes of unknown function whose patterns of expression and regulation suggest that they could be involved in sporulation. Disruption of two such newly identified genes, *yabP* and *yabQ*, blocked sporulation at a late stage.

hidden Markov models | functional genomics | DNA arrays

Sporulation by *Bacillus subtilis* is governed by a program of gene transcription involving more than 125 genes (reviewed in ref. 1). However, a description of global changes in gene expression during the course of spore formation has been lacking. Recent methods for creating DNA arrays with large numbers of individual genes has made it possible to visualize genome-wide changes in gene transcript levels, as, for example, during sporulation in budding yeast and during the response of human fibroblasts to serum (2, 3). Taking advantage of the availability of a complete genome sequence for *B. subtilis* (4), we sought to apply transcriptional profiling to the process of sporulation in this bacterium.

Entry into sporulation is principally governed by the DNA-binding protein Spo0A, which is both an activator and a repressor of transcription. The activity of Spo0A, which is a member of the response regulator family of proteins, is controlled through phosphorylation by a phosphorelay that integrates environmental and physiological signals in the decision to sporulate (5). Among the targets of the phosphorylated form, Spo0A~P, are one or more unknown genes involved in the formation of a polar septum, which divides the cell into forespore and the mother cell compartments (6). Spo0A~P also directs the transcription of genes and operons (*spoIIA*, *spoIIE*, *spoIIG*) involved in the activation of the compartment-specific regulatory proteins σ^F and σ^E, which direct transcription in the forespore and the mother cell, respectively (1). Spo0A also activates transcription of certain genes indirectly by repressing the gene for the "transition-state regulator" AbrB (7). Among the targets of AbrB is the gene (*sigH*) for the early-appearing sporulation transcription factor σ^H (8). Later in development, σ^F is replaced in the forespore by σ^G and σ^E is replaced in the mother cell by σ^K (1).

We sought to identify genes that are activated or repressed at early to middle stages of sporulation by comparing transcripts present during growth and sporulation of the wild type with transcripts present in mutants for Spo0A and σ^F. Because σ^F is required for the appearance of σ^E and all other transcription factors downstream in the sporulation pathway (1), genes whose expression depends on Spo0A, but not on σ^F, would be candidates for targets of Spo0A. Genes whose expression depends on both Spo0A and σ^F could be under the direct control of σ^F or of a downstream transcription factor in the sporulation pathway. Two additional approaches were used to identify genes under the control of Spo0A and other regulatory proteins. In one, we monitored gene expression in cells engineered to produce a constitutively active form of Spo0A (Spo0A-Sad67) during growth (9). In a second approach, we used hidden Markov models (HMMs) trained on known promoter sequences to identify genes that were likely to be under the control of Spo0A, σ^F, or σ^E. This analysis enabled us to discover many additional genes that are likely to be part of the sporulation program, two of which are shown to be essential for spore formation.

Materials and Methods

Strains and Media. The strains used were as follows: RL2242 (*spo0A*::*spc*); RL1104 (*amyE*::P_{spac}-*spo0A-sad67 cat*, *spo0A*Δ*erm*); RL1265 (*sigF*::*kan*); PY79 (prototroph); RL2243 (*yabP*::*spc*); RL2244 (*yabQ*::*tet*); RL2245 [*yabP*::*spc thrC*::*pCW63*(*cotD-lacZ erm*)]; and RL2246 [*yabQ*::*tet thrC*::*pCW63*(*cotD-lacZ erm*)]. Cells were grown overnight on LB agar and inoculated into 30 ml of Difco sporulation medium (DSM) at $OD_{600} \approx 0.05$ (10). Parallel cultures of PY79, RL2198, and RL1265 were grown at 37°C until they reached an OD_{600} of ≈0.3, and samples were then diluted into 30 ml of fresh DSM to achieve similar cell densities. At times $T_{-1.5}$, T_0, and T_2 relative to the end of the exponential growth, 7 ml of culture was withdrawn, and the cells were pelleted and frozen in liquid N_2. Induction of P_{spac}-*spo0A-sad67* was accomplished by growing RL1104 in LB at 37°C until $OD_{600} \approx 0.3$. The culture was split, and isopropyl β-D-thiogalactoside (IPTG) was added to one half at a final concentration of 1 mM. Cells were harvested at 0, 15, 30, and 60 min, and handled as before. Deletion mutants of *yabP* (RL2243) and *yabQ* (RL2246) were created by using the long-flanking homology PCR strategy (11).

Transcriptional Profiling. We used custom nylon-substrate PCR-product-based arrays. PCR amplification in 96-well plates was carried out with a set of primers for 4,100 *B. subtilis* ORFs (Eurogentech; Seraing, Belgium). PCR products were subjected

Abbreviations: HMM, hidden Markov model; IPTG, isopropyl β-D-thiogalactoside.

§Present address: Dept. of Biochemistry, Stanford University, Stanford, CA 94305.

¶To whom reprint requests should be addressed at: Department of Molecular and Cellular Biology, The Harvard Biological Laboratories, 16 Divinity Ave., Cambridge, MA 02138. E-mail: losick@biosun.harvard.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.140209597. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.140209597

to electrophoresis on agarose gels and scored, and ≈96% of the 4,100 ORFs of *B. subtilis* were represented. RNA was prepared by using the hot acid/phenol method (protocols available at http://mcb.harvard.edu/losick). Random hexamer-primed reverse transcription reactions in the presence of [α-^{33}P]dCTP using 7.5 μg of total RNA were used to create cDNA probes. Spin-purified probes were divided and hybridized to replicate array membranes at 68°C for 18 h. Membranes were washed, baked, and exposed overnight to phosphoimager screens. A Fuji phosphoimager and ARRAYVISION software (Imaging Research, St. Catherines, ON, Canada) were used to quantify hybridization intensities, and for background subtraction. Spots were associated with ORFs by using EXPRESSION EXPLORER (Millennium Pharmaceuticals), and intensity distributions were median normalized to a value of 1. A minimum threshold value of 0.1 was set for all data points to avoid spurious expression level ratios at the bottom of the spot intensity range. Datasets were imported into a FILEMAKER PRO (FileMaker) database (http://mcb.harvard.edu/losick). Expression level ratios were calculated with averaged values obtained from replicate filters, and we considered only those ORFs showing at least a 3-fold difference. We superimposed on this basic criterion a statistical heuristic to eliminate ORFs that gave inconsistent hybridization by calculating a 90% confidence interval on the replicate data and asking that ratios calculated at the extremes of the confidence intervals remained greater than 3-fold.

HMMs. Multiple alignments of known promoter recognition sequences for each specific transcription factor were annotated with hand-specified model structures; regions spanning the −35 to −10 of each promoter were assigned to match states, and bases in the spacer region were replaced with ambiguity characters representing the background nucleotide distribution of *B. subtilis*. Separate models were built in both orientations, and for known allowed sizes of the spacer. The HMMBUILD module of the HMMER 2.1.1 package (http://hmmer.wustl.edu) was used to create HMMs global with respect to the model, but local with respect to the search string (12). To model Spo0A-activated σ^A-dependent promoters, we made hybrid HMMs that combined the probabilities of known "−35" regions for promoters recognized by σ^A–RNA polymerase with the derived probabilities of Spo0A-binding sites, separated by spacers the same size as for known Spo0A-activated σ^A-dependent promoters (*spoIIG*, 21 bp; *spoIIE*, 22 bp). In the case of the σ^F model, little training data are available, but since promoter-recognition sequences for σ^F and σ^G have significant similarities, we also constructed a model trained on known σ^F and σ^G promoters. The HMMSEARCH module was then used to search genome release 14.2 (4, 13). Custom software was used to correlate the position of hits with respect to nearby ORFs.

Results and Discussion

Transcriptional Profiling During Sporulation. Experiments were carried out with RNA from wild-type cells and from cells mutant for Spo0A or for σ^F. Cells were harvested 1.5 h before ($T_{-1.5}$), at (T_0), and 2 h after (T_2) the start of sporulation. Radioactive cDNAs were generated from the RNAs and incubated with nylon filters spotted with PCR products corresponding to almost all of the annotated ORFs in the genome. As an indication of reproducibility, spot intensities were compared between pairs of filters hybridized with the same radioactive cDNAs. The results (Fig. 1*A* is a representative example) showed good reproducibility between filters; in all cases the Pearson correlation coefficient (r) of replicate filter pairs was >0.96. In contrast, a scatterplot (Fig. 1*B*) comparing the intensity of spots generated with cDNAs from wild-type cells with those generated with cDNAs from a Spo0A mutant showed that the mutation caused significant changes in the expression profile. As an example of

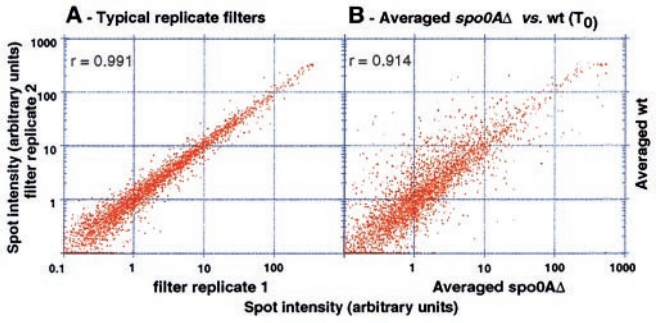

Fig. 1. Logarithmic-scale plots of normalized spot intensities (arbitrary units). (*A*) Typical replicate filters are shown with the intensity of each spot plotted versus the equivalent spot in a replicate filter. (*B*) Averaged spot intensities from a filter hybridized with probe from wild-type cells at T_0 versus averaged spot intensities from a filter hybridized with probe from the Spo0A mutant at T_0. Each graph represents 4,100 individual spots. r is the Pearson correlation coefficient.

the primary data, Fig. 2 shows a close-up of the autoradiographs for a small region of arrays that had been hybridized with radioactive cDNA generated from RNAs from wild-type and σ^F mutant cells undergoing sporulation.

We used the clustering technique of Eisen *et al.* (14) to represent the expression profiles of all genes (586) whose transcript levels were at least 3-fold different between the wild type and the Spo0A mutant during at least one of the time points (Fig. 3*A*). Of these 586 genes, 266 exhibited at least a 5-fold dependence in transcript levels on Spo0A, and 83 exhibited greater than a 10-fold dependence. The first three columns of the cluster diagram compare the ratio of the hybridization signals between the wild type and the Spo0A mutant for RNAs from cells at $T_{-1.5}$, T_0, and T_2. The fourth column compares the ratio of signals between the wild type and the σ^F mutant at T_2.

Fig. 3*A* shows that the genes fell into three categories. Category I were genes whose expression was higher in the wild type than in the Spo0A mutant (during at least one of the time points) but whose expression was not significantly different between wild type and the σ^F mutant. Category I therefore represented genes whose expression was dependent on Spo0A but not on σ^F. Category II was genes with lower expression in the wild type than in the Spo0A mutant during at least one of the time points. This category represented genes whose expression was inhibited by Spo0A. Finally, category III was genes whose expression was higher in the wild type than in both the Spo0A mutant and the σ^F mutant. This category represented genes whose expression was under the control of σ^F or of a downstream transcription factor in sporulation.

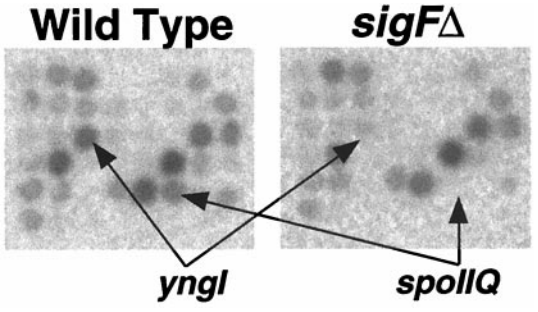

Fig. 2. Close-ups of autoradiographs of filter arrays probed with cDNA from wild-type (left) and σ^F (*sigF*) mutant cells (right) at T_2. The arrows point to the known σ^F-controlled gene *spoIIQ* and to a gene (*yngI*) that was not previously known to be under sporulation control.

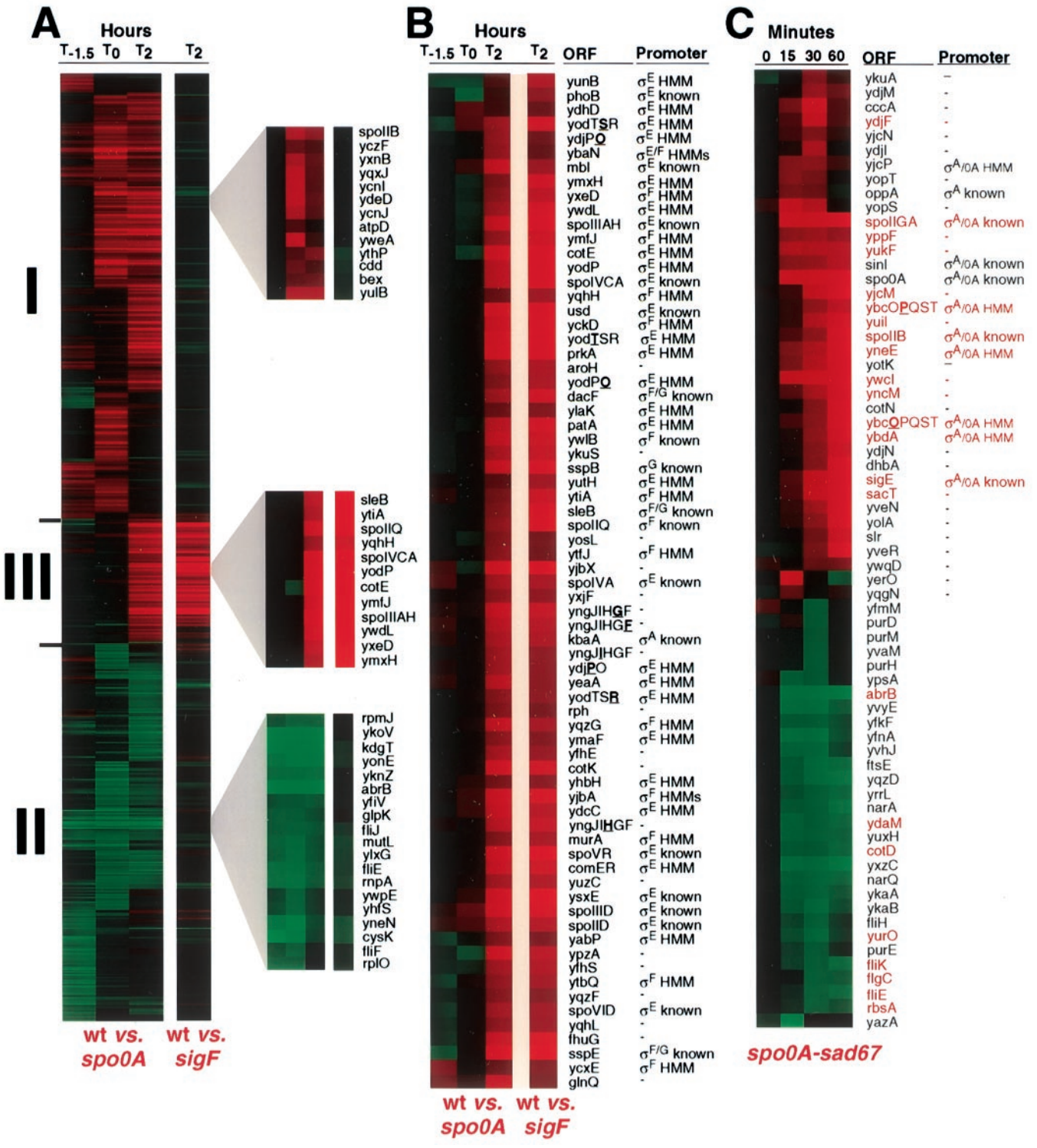

Fig. 3. Eisen plots and HMM predictions. (*A*) Eisen plot indicating the profiles of 586 genes whose level of expression depended on Spo0A. The genes were ordered so that those with similar expression patterns were grouped together. Columns 1–3 display ratios of normalized spot intensities of hybridization signals for probe from the wild type versus probe for the Spo0A (*spo0A*) mutant at the indicated times. Column 4 displays the ratio of the hybridization signals for the wild type versus the σ^F (*sigF*) mutant at T_2. The zoom boxes on the right show representative genes from each of the classes (see text). (*B*) An annotated Eisen plot is shown for the 66 genes whose expression was higher in the wild type than in both the Spo0A mutant and the σ^F mutant at T_2. When available, known promoters or HMM predictions are indicated. (*C*) An annotated Eisen plot for the 67 genes that were differentially expressed by 15 or 30 min after the addition of IPTG during vegetative growth to cells of a strain harboring P_{spac}-*spo0A-sad67*. The plot compares transcript levels between cells treated with IPTG and untreated cells. Promoters or HMM predictions are indicated for genes induced in response to IPTG. Genes whose expression was dependent on Spo0A but not σ^F in *A* and were induced in response to IPTG in this experiment are indicated by the use of red text. Genes that are thought to lie in an operon with other genes in the same expression category are indicated by boldfacing and underlining. Hybridization ratios are displayed colorimetrically, with stronger hybridization for the wild type compared with the mutant shown as shades of red (ratio >1) and stronger hybridization for the mutant (ratio <1) shown as shades of green (*A* and *B*). Likewise, red and green indicate higher and lower ratios, respectively, for IPTG-treated cells relative to untreated cells (*C*).

Genes Whose Expression Depended on Spo0A but Not σ^F. The largest group of genes (283) were in category I. Representative members are shown in the zoom in Fig. 3*A*. Consistent with the idea that genes in this category are directly or indirectly dependent on Spo0A~P for their expression, only 47 of the 283 genes in category I showed a significant dependence on *spo0A* during the mid-exponential phase of growth (the $T_{-1.5}$ time point); it was not until at or after the start of sporulation, when Spo0A~P becomes abundant, that the majority of genes in this category showed a clear difference in expression levels between the wild type and the *spo0A* mutant.

Category I included sporulation genes whose transcription was previously known to be under the direct control of Spo0A, such as *spo0F, spoIIA,* and *spoIIG* (1, 15). It also included sporulation genes whose transcription was known to be indirectly dependent on Spo0A, such as *kinA* and *spoVG* (16, 17). Included in category I were various stationary-phase genes, including *aprE*, *nprB*, *vpr*, and *epr*, which encode proteases, *pel*, which encodes a pectate lyase (18), and *csn*, which specifies a chitosanase. Also in category I were genes for a nonribosomal peptide synthase (*srfAB, srfAD*), a peptide antibiotic (*sbo*), and polyketide synthetases (*pksD* and *pksE*). Several metabolic genes were also in this category, including genes involved in glutamine uptake (*glnH*) and histidine utilization (*hutH*, *hutI*, *hutG*).

In addition, category I included an operon (*ctaC,D,E,F,G*) encoding an alternative terminal cytochrome *c* oxidase of the caa_3 family. This is of interest because sporulation defects arise from mutations in genes involved with energy production, notably those in the tricarboxylic acid cycle (19). A sporulation phenotype is also seen in a mutant of the divergently transcribed *ctaA* gene, which is needed for the biosynthesis of heme A, a prosthetic group required by the caa_3 oxidases (20, 21). Significant expression of *ctaC,D,E,F,G* occurs only during stationary phase, and it is subject to catabolite control (22). If the encoded cytochrome oxidase is involved with sporulation, its regulation could contribute to the catabolite control of sporulation.

Other notable genes in this category were those involved in oligopeptide transport (*dppB, dppD, dppE, appD, appF, appB*), genes for peptide pheromones (*phrE, phrF, phrG, phrI*), and genes (*rapC, rapE, rapF, rapI)* for regulatory aspartyl phosphatases, which are the presumptive targets of the pheromones (23). Many of the previously known genes in category I are indirectly linked to Spo0A. For example, some of the extracellular protease genes are negatively regulated by AbrB, the gene for which is negatively regulated by Spo0A~P or by the product of *hpr*, which is activated by AbrB (24).

The large majority of genes in category I were of unknown function. Many either exhibited no significant homology to other genes in the databases or were homologous to other proteins of unknown function. Some of the previously unknown genes in category I did exhibit homology to genes of known function. For example, *ykuA* encodes a protein with similarity to penicillin-binding proteins. Several of the genes of unknown function in this category appear to be organized in operons. Two examples are the gene cluster that includes *yxbB*, *yxbA*, and *yxnB* and the cluster that includes *ybcO, ybcP, ybcQ, ybcS, ybcT, ybdA,* and *ybdB*.

Genes Whose Expression Was Inhibited by Spo0A. Category II represented genes (242) whose expression was directly or indirectly inhibited by Spo0A. Representative members of category II are shown in the zoom in the bottom right of Fig. 3*A*. Category II included genes that were previously known to be subject to repression by Spo0A, such as the transition state regulator *abrB* (7). As expected, we found that *abrB* transcript levels were much higher in the Spo0A mutant than in the wild type at all of the time points. Category II also included genes whose expression was indirectly linked to Spo0A. For example, it included the ribose transport operon (*rbsR,K,D,A,C,B*), whose expression under conditions of catabolite repression is known to be dependent on AbrB (25). Thus, expression of the ribose transport operon is enhanced in a Spo0A mutant because of overexpression of *abrB*. Other notable genes in category II were chemotaxis genes (*mcpA*, *fliJ*, *cheB*, *cheA*, *cheW*, and *cheC*), motility genes (*flgC*, *flgE*, *flhO*, *fliD*, *fliF*, *fliG*, *fliK*, and *fliY*) and autolysin genes (*lytD* and *lytE*), which were overexpressed in the *spo0A* mutant. The presence of chemotaxis, motility, and autolysin genes in this category illustrates the complex pleiotropy of *spo0A* mutations. These genes are transcribed by the σ^D transcription factor, which is encoded by *sigD*, and expression of *sigD* is positively regulated by the SinR transition state regulator (26). However, SinR activity is impaired when it interacts with SinI (27), the gene for which is transcribed in a Spo0A~P-dependent manner. A *spo0A* mutation therefore results in abnormally high levels of σ^D-dependent gene transcription.

Genes Whose Expression Depended on Both Spo0A and σ^F. Category III consisted of 66 genes, which are shown in the middle zoom on the right of Fig. 3*A* and in Fig. 3*B*. Consistent with the idea that expression of these genes was dependent on σ^F, genes in category III were expressed more highly in the wild type than in either the Spo0A mutant or the σ^F mutant, and in almost all cases this selective expression was not observed until T_2, the earliest time point by which σ^F would have become active in directing sporulation gene transcription. Because the appearance of σ^E requires the prior activation of σ^F (1) and because some σ^E-directed transcription would have commenced by T_2, category III was expected to include genes under the control of σ^F or σ^E.

Category III included many genes known to be under the control of σ^F (or both σ^F and σ^G), such as *spoIIQ*, *dacF*, *sleB, sspE,* and *ywlB*, or σ^E, such as *cotE*, *spoIVA*, *spoIVCA*, *spoIID*, *spoIIID*, *spoIIIAH*, *spoVR*, *spoVID*, *usd*, *mbl,* and *ysxE*. However, certain other genes known to be under the control of σ^F, such as *spoIIR* and *katX,* or σ^E, such as the *spoIIIA* operon and *spoIVCB*, were missing from category III because of the low level and/or timing of their expression. Category III included many genes not known to be expressed during sporulation. Some of these newly identified sporulation-controlled genes encode proteins with homology to proteins with known functions. For example, we detected an apparent operon (*yngJ,I,H,G,F,E*) whose products appear to be involved in the metabolism of long-chain fatty acids. Thus, the products of *yngJ*, *yngI*, *yngH*, *yngG*, *yngF*, and *yngE* are similar, respectively, to a butyryl-CoA dehydrogenase, a long-chain acyl-CoA synthetase, a biotin carboxylase, a hydroxymethylglutaryl-CoA ligase, a hydroxybutyryl-CoA dehydratase, and a propionyl-CoA carboxylase. Likewise, *yodT, yodS,* and *yodR* appear to constitute an operon whose products are similar, respectively, to an adenosylmethionine-8-amino-7-oxononanoate aminotransferase, a 3-oxoadipate CoA-transferase, and a butyrate-acetoacetate CoA-transferase. In keeping with a sporulation role for fatty acid metabolism, *mmg*, a previously identified operon involved with fatty acid utilization, is known to exhibit a σ^E-dependent mode of transcription (28).

Category III also included several previously annotated genes that had not been recognized as being expressed under sporulation control. Examples are *murA*, which is homologous to an *Escherichia coli* gene involved in peptidoglycan biosynthesis (29), *prkA*, which encodes a serine protein kinase of unknown function (30), and *comER* (31).

Alterations in the Global Pattern of Gene Expression in Cells Engineered to Produce a Constitutively Active Form of Spo0A During Growth. As a complementary approach to identifying genes under the control of Spo0A, we examined the effect on global gene expression of producing a constitutively active form of Spo0A during growth. Spo0A-Sad67 is a mutant form of Spo0A

whose activity does not require phosphorylation (9). Synthesis of Spo0A-Sad67 was induced during growth by using the inducer IPTG and a fusion of *spo0A-sad67* to the IPTG-inducible promoter P_{spac}. RNA was extracted from cells collected at 0, 15, 30, and 60 min after the addition of inducer and from untreated cells collected at the same times. Transcriptional profiling revealed 29 genes for which there was at least a 3-fold difference in transcript levels between the treated and the untreated cells by 15 min after the addition of inducer. By 30 min this number increased to 67 and, finally, by 60 min, 246 genes were observed whose transcript levels were significantly altered by treatment with inducer. Genes whose expression was altered in the earlier time points were more likely to be under the direct control of Spo0A than genes for which an alteration in transcript levels was not observed until an hour after the addition of inducer.

Among genes whose expression rapidly increased in response to IPTG were known targets of Spo0A, such as *spoIIGA*, *sigE*, *spoIIB*, and *sinI.* These genes are transcribed from Spo0A-controlled promoters that are recognized by σ^A–RNA polymerase (1). For example, after 15 min the level of *spoIIGA* transcripts was 25-fold higher in the treated than in the untreated cells. Conversely, transcript levels for *abrB*, which is repressed by Spo0A (7), were 26-fold lower in the induced cells than in the uninduced cells at the same time point. Of the 67 genes whose expression was altered by the 30-min time point, 37 genes had higher expression in the treated than in the untreated cells. Of these, 26 were previously uncharacterized genes, and several had similarities to known genes: *ykuA* is similar to penicillin-binding gene 2b of *Streptococcus pneumonia*; *ywqD*, *yveR,* and *yveN* are similar to spore coat or capsular polysaccharide biosynthetic genes (*ywqD* to *capB* of *Staphylococcus aureus*; *yveR* to numerous galactosyl- and glycosyltransferases, including the *spsA* spore coat polysaccharide biosynthesis gene of *Synechocystis* sp.; *yveN* to lipopolysaccharide synthesis genes, including *wbnE* of *E. coli*); and *yvhJ*, *yerO,* and *slr* are similar to transcriptional regulators (*yvhJ* to *lytR* of *B. subtilis*; *yerO* to members of the *tetR/acrR* family, and *slr* to *sinR* of *B. subtilis*). Also noteworthy was an apparent operon formed by *ywcI* and *sacT*, the antiterminator of a sucrose utilization operon.

Among the genes whose transcript levels decreased in response to inducer were several flagellar protein genes (*fliH, fliK, flgC, fliE*), an uncharacterized (in *B. subtilis*) homolog of the *E. coli* cell division gene *ftsE*, and several genes involved in purine biosynthesis (*purD, purE, purH, purM*). A possible role of Spo0A in the repression of purine biosynthetic genes is intriguing because it has long been appreciated that the onset of sporulation is accompanied by a decrease in the cellular concentrations of GDP and GTP, and inhibition of guanine synthesis can induce sporulation in rich medium (32). Conceivably, Spo0A sets up a self-reinforcing loop for entry into sporulation by repressing purine biosynthetic genes.

Finally, we note that among the genes whose expression was enhanced or repressed in response to IPTG were 24 genes whose expression was influenced by Spo0A during sporulation in the transcriptional profiling experiment of Fig. 3*A*. These 24 genes, which are annotated with red text in Fig. 3*C*, are our best candidates for genes that are directly under the control of Spo0A. Included in this group is one anomalous case [*cotD* (1)], which we presume is attributable to a misamplified PCR product.

Promoter Recognition Sequence Prediction. To gain further insight into the regulatory mechanisms governing the expression of the genes identified in our transcriptional profiling experiments, we constructed a series of HMMs that probabilistically modeled promoter sequences recognized by RNA polymerase containing σ^F and σ^E, as well as Spo0A-activated promoters recognized by σ^A–RNA polymerase (12). Fig. 3*B* shows that of the 48 novel or poorly characterized genes that were found to have transcription profiles characteristic of regulation by σ^F or σ^E, we were able to predict a likely promoter for 34, of which 11 were assigned to control by σ^F, and 22 were assigned to control by σ^E (one gene, *ybaN*, was predicted to have both σ^E- and σ^F-recognized promoters). Although there are doubtless false positives in our assignments, the percentage (70%) of promoters that could be assigned to control by σ^F or σ^E was higher than could occur by chance, and the promoters predicted for the genes identified by transcriptional profiling tended to be located within 100 bp of the initiation codon.

Fig. 3*C* presents the results of the HMM analysis for Spo0A-activated promoters recognized by σ^A–RNA polymerase. Among genes identified as potential targets of Spo0A through the use Spo0A-Sad67 (Fig. 3*C*), *yjcP*, *yneE, ybdA,* and *ybcO,P* were preceded by sequences that significantly conformed to our model for Spo0A-activated promoters. Of particular interest is a cluster of seven genes consisting of the genes *ybcO, ybcP, ybcQ, ybcS,* and *ybcT* and *ybdA* and *ybdB*, which could constitute an operon. This operon is a strong candidate for a transcription unit under the direct control of Spo0A: it was activated in a Spo0A-dependent manner during sporulation; it was switched on during growth in response to the induced synthesis of Spo0A-Sad67 (Fig. 3*B*); and it appears to be preceded by a sequence conforming to that of Spo0A-activated promoters recognized by σ^A–RNA polymerase. None of the products encoded by this putative operon are similar to previously characterized proteins, except for the products of *ybdA* and *ybdB*, which resemble, respectively, the binding and permease components of ABC transporters.

***yabP* and *yabQ* Are Required for Sporulation.** Our profiling and HMM analyses have revealed many previously uncharacter-

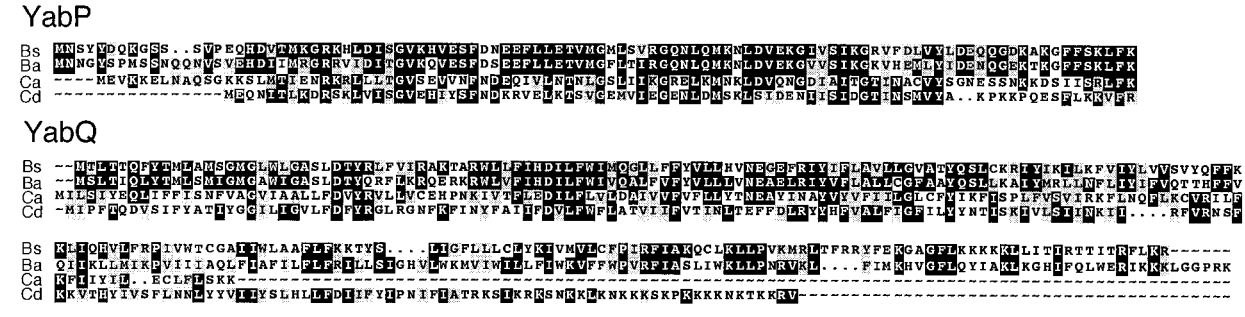

Fig. 4. Inferred amino acid sequences for YabP and YabQ from *B. subtilis* (Bs) and orthologs (deduced by GeneMark and related algorithms; ref. 33) from *B. anthracis* (Ba), *C. acetobutylicum* (Ca), and *C. difficile* (Cd). *B. anthracis*, *C. acetobutylicum,* and *C. difficile* orthologs were identified in unannotated data from unfinished sequencing projects (http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html).

ized genes that appear to be under sporulation control. We have begun to build null mutations of many of these genes in an effort to identify those that play a role in sporulation. One such case is that of the adjacent *yabP* and *yabQ* genes, which appear to constitute an operon. The upstream member of the putative operon, *yabP,* had an expression profile consistent with regulation by σ^F or σ^E. HMM analysis predicted a sequence, beginning 72 bp upstream of the initiation codon, with similarity to promoter sequences recognized by σ^E-containing RNA polymerase. *yabP* is predicted to encode a 100-residue protein, and *yabQ* is predicted to encode a 211-residue protein with five membrane-spanning domains. The putative *yabPQ* operon is immediately upstream of the cell division gene *divIC* and is located near and between known sporulation genes *spoVT* and *spoIIE*. We found that the replacement of *yabP* or *yabQ* with a drug-resistance cassette led to a severe sporulation defect. Similar conclusions have been reached by the Japanese functional analysis network (http://bacillus.genome.ad.jp). Examination of sporulating cells of the *yabP* and the *yabQ* mutants by phase-contrast microscopy revealed the production of phase-gray spores that failed to brighten, indicating a block at a late stage of sporulation (data not shown). We found orthologs of both *yabP* and *yabQ* in *Clostridium acetobutylicum* ATCC824 (Genome Therapeutics Corp.), *Clostridium difficile* (Sanger Centre), and *Bacillus anthracis* (The Institute for Genomic Research) (an alignment is shown in Fig. 4). In these organisms, the local gene order is the same as in *B. subtilis* (*yobO-yabP-yabQ-divIC*), but the same region in *Listeria monocytogenes*, a non-spore-forming Gram-positive bacterium, is simply *yobO-divIC* (as brought to our attention by P. Stragier, personal communication).

Summary. Transcriptional profiling reinforces the view that sporulation involves global changes in gene expression and has led to the identification of many previously uncharacterized genes whose expression is under sporulation control. In total, we identified 586 genes, representing more than 10% of the ORFs in the *B. subtilis* genome (4), whose level of expression was altered relative to that of a Spo0A mutant at least 3-fold during the early to middle stages of sporulation. We were able to tentatively assign many of these genes to regulatory classes through the use of mutants for the sporulation regulatory proteins Spo0A and σ^F, through the use of cells engineered to produce Spo0A-Sad67 during growth, and through an informatics approach based on HMMs trained on known sporulation promoters.

The large majority of genes identified in our analysis, corresponding to categories I and III, were activated or repressed early in sporulation under the direct or indirect control of Spo0A. This indicates that the master regulator for entry into sporulation causes profound changes in the global pattern of gene expression as cells switch from growth to differentiation. Among the newly identified candidates for genes under the direct control of Spo0A are *yjcP*, *yneE,* and a putative seven-member operon consisting of the genes *ybcO, ybcP, ybcQ, ybcS, ybcT*, *ybdA,* and *ybdB*. Transcriptional profiling and informatics also led to the identification of several previously uncharacterized genes that appear to be under the control of σ^F or σ^E. Two of these genes, *yabP* and *yabQ*, were shown to be essential for spore formation, and we suspect that others in this category play a role in sporulation. Disruption of previously uncharacterized genes from our analysis should provide a more complete view of how global changes in gene expression help govern differentiation.

We thank N. Su and C. Murphy for assistance with profiling and R. Britton, E. Gonzalez, F. Gueiros-Filho, C. Price, and A. L. Sonenshein for advice. Work at Harvard was supported by National Institutes of Health Grant GM18568 to R.L. and a gift from Millennium Pharmaceuticals. P.E. was a postdoctoral fellow of the Human Frontier Science Program.

1. Stragier, P. & Losick, R. (1996) *Annu. Rev. Genet.* **30,** 297–241.
2. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. & Herskowitz, I. (1998) *Science* **282,** 699–705.
3. Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., *et al.* (1999) *Science* **283,** 83–87.
4. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., *et al*. (1997) *Nature (London)* **390,** 249–256.
5. Burbulys, D., Trach, K. A. & Hoch, J. A. (1991) *Cell* **64,** 545–552.
6. Levin, P. A. & Losick, R. (1996) *Genes Dev.* **10,** 478–488.
7. Strauch, M., Webb, V., Spiegelman, G. & Hoch, J. A. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 1801–1805.
8. Weir, J., Predich, M., Dubnau, E., Nair, G. & Smith, I. (1991) *J. Bacteriol.* **173,** 521–529.
9. Ireton, K., Rudner, D. Z., Siranosian, K. J. & Grossman, A. D. (1993) *Genes Dev.* **7,** 283–294.
10. Harwood, C. R. & Cutting, S. M. (1990) *Molecular Biological Methods for Bacillus* (Wiley, New York).
11. Wach, A. (1996) *Yeast* **12,** 259–265.
12. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ. Press, Cambridge, U.K.).
13. Moszer, I., Glaser, P. & Danchin, A. (1995) *Microbiology* **141,** 261–268.
14. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.
15. Asayama, M., Yamamoto, A. & Kobayashi, Y. (1995) *J. Mol. Biol.* **250,** 11–23.
16. Predich, M., Nair, G. & Smith, I. (1992) *J. Bacteriol.* **174,** 2771–2778.
17. Zuber, P. & Losick, R. (1987) *J. Bacteriol.* **169,** 2223–2230.
18. Nasser, W., Awade, A. C., Reverchon, S. & Robert-Baudouy, J. (1993) *FEBS Lett.* **335,** 319–326.
19. Ireton, K., Jin, S., Grossman, A. D. & Sonenshein, A. L. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 2845–2849.
20. Mueller, J. P. & Taber, H. W. (1989) *J. Bacteriol.* **171,** 4967–4978.
21. Svensson, B., Lubben, M. & Hederstedt, L. (1993) *Mol. Microbiol.* **10,** 193–201.
22. Liu, X. & Taber, H. W. (1998) *J. Bacteriol.* **180,** 6154–6163.
23. Perego, M. (1998) *Trends Microbiol* **6,** 366–370.
24. Perego, M. & Hoch, J. A. (1988) *J. Bacteriol.* **170,** 2560–2567.
25. Strauch, M. A. (1995) *J. Bacteriol.* **177,** 6727–6731.
26. Rashid, M. H. & Sekiguchi, J. (1996) *J. Bacteriol.* **178,** 6640–6643.
27. Bai, U., Mandic-Mulec, I. & Smith, I. (1993) *Genes Dev.* **7,** 139–148.
28. Bryan, E. M., Beall, B. W. & Moran, C. P., Jr. (1996) *J. Bacteriol.* **178,** 4778–4786.
29. Marquardt, J. L., Siegele, D. A., Kolter, R. & Walsh, C. T. (1992) *J. Bacteriol.* **174,** 5748–5752.
30. Fischer, C., Geourjon, C., Bourson, C. & Deutscher, J. (1996) *Gene* **168,** 55–60.
31. Inamine, G. S. & Dubnau, D. (1995) *J. Bacteriol.* **177,** 3045–3051.
32. Lopez, J. M., Marks, C. L. & Freese, E. (1979) *Biochim. Biophys. Acta* **587,** 238–252.
33. Besemer, J. & Borodovsky, M. (1999) *Nucleic Acids Res.* **27,** 3911–3920.