**Bacillus subtilis** Systems Biology: Applications of -Omics Techniques to the Study of Endospore Formation

ASHLEY R. BATE, RICHARD BONNEAU, and PATRICK EICHENBERGER

Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003

**ABSTRACT** Endospore-forming bacteria, with **Bacillus subtilis** being the prevalent model organism, belong to the phylum Firmicutes. Although the last common ancestor of all Firmicutes is likely to have been an endospore-forming species, not every lineage in the phylum has maintained the ability to produce endospores (hereafter, spores). In 1997, the release of the full genome sequence for **B. subtilis** strain 168 marked the beginning of the genomic era for the study of spore formation (sporulation). In this original genome sequence, 139 of the 4,100 protein-coding genes were annotated as sporulation genes. By the time a revised genome sequence with updated annotations was published in 2009, that number had increased significantly, especially since transcriptional profiling studies (transcriptomics) led to the identification of several genes expressed under the control of known sporulation transcription factors. Over the past decade, genome sequences for multiple spore-forming species have been released (including several strains in the **Bacillus anthracis**/**Bacillus cereus** group and many **Clostridium** species), and phylogenomic analyses have revealed many conserved sporulation genes. Parallel advances in transcriptomics led to the identification of small untranslated regulatory RNAs (sRNAs), including some that are expressed during sporulation. An extended array of -omics techniques, i.e., techniques designed to probe gene function on a genome-wide scale, such as proteomics, metabolomics, and high-throughput protein localization studies, have been implemented in microbiology. Combined with the use of new computational methods for predicting gene function and inferring regulatory relationships on a global scale, these -omics approaches are uncovering novel information about sporulation and a variety of other bacterial cell processes.

**THE BACILLUS SUBTILIS GENOME SEQUENCE AND GENOMICS OF SPORULATION**

**B. subtilis** Genomics

The principal **B. subtilis** laboratory strain, strain 168, is derived from a parent strain isolated in Marburg, Germany, following a mutagenesis procedure (1). The popularity of this strain arose after it was shown to be competent for genetic transformation (2, 3), which paved the way for myriad molecular genetics analyses that led to a detailed understanding of the biology of **B. subtilis** and related Gram-positive bacteria. It is therefore not surprising that strain 168 was the first Gram-positive species to have its entire genome sequenced, at a time when sequencing was a laborious and expensive process. The project to sequence the genome was set up in 1987 by a consortium of over 30 laboratories and took about 10 years to complete. Each laboratory was assigned a
Bate et al.

different region of the chromosome and used their own cloning and sequencing strategies to manage their assigned portion of the genome (4). The final genome sequence contained 4,214,810 base pairs, and the original annotation included 4,100 protein-coding genes (5). Following the development of sequencing technologies that were considerably faster and more efficient, the genome of B. subtilis strain 168 was resequenced and cleared of sequencing errors in 2009 (6). The most recent update of the annotation brought the total of protein-coding genes to 4,458 (7).

The improved ease and expanded access of sequencing technologies also shed light on the ancestry of several other B. subtilis strains. Although strain 168 is a popular model organism, it has limitations that have led numerous laboratories to study other strains. Through a process of domestication in the laboratory, strain 168 and derivatives lost swarming motility (8) and the ability to construct robust biofilms (2). Therefore, to study these processes, researchers have turned to the undomesticated NCIB3610 strain, which is considered to be an authentic representation of the B. subtilis Marburg parent strain. Furthermore, due to the presence of a trpC2 allele, strain 168 is a tryptophan auxotroph. Thus, in order to generate Trp+ derivatives of strain 168, a trpC+ allele had to be provided by a donor strain, usually strain W23, a threonine auxotroph that appears to have arisen from an independent isolate (2). A study analyzing the genomic region surrounding trpC for 17 of the B. subtilis legacy strains has shown that a number of strain 168 Trp+ derivatives have W23-parent DNA, implying that many B. subtilis strains used by several laboratories around the world are mosaics of strains 168 and W23 (10). The determination of strain ancestry is particularly valuable considering that some of the -omics work performed in B. subtilis has been done with strain PY79 (11), which has the greatest divergence from the original strain among the 168 legacy strains (10). It should be emphasized that the genome size of strain PY79 is smaller (by 180 kb) than that of strain 168, since it contains four large deletions, in particular the regions corresponding to the SPβ prophage (11) and the conjugative transposon ICEB1 (12).

B. subtilis W23 shares a 3.6-Mb core genome with strain 168 (13). Strains W23 and 168 are fundamentally different in their cell wall composition, with teichoic acid polymers of strain W23 resembling those of Staphylococcus aureus (14, 15). In addition, the sporulation killing factor (SKF) and competence proteins ComQXP, two characteristic traits of strain 168, are not found in strain W23 (13). An increasing number of B. subtilis strains and closely related species have been sequenced recently, and analyses of these genome sequences revealed that despite a high degree of conservation, each strain harbors numerous unique regions (16–18). Most strain-specific differences appear to originate with the transfer of genetic elements including phages, plasmids, and insertion sequences. Many differences affect secondary metabolism and developmental programs (such as sporulation and competence), suggesting that these pathways are plastic and may be under greater environmental selection than other parts of the cellular machinery. This information is especially valuable to understand how B. subtilis can occupy the variety of environmental niches in which it has been shown to reside (see Fajardo-Cavazos et al. [105] and Mandic-Mulec et al. [106]).

Comparative Genomics of Sporulation

In addition to B. subtilis, the genomes of many other sporeformers have been sequenced, including many from the family Bacillaceae and from the order Clostridiales. Of particular note are several species in the Bacillus anthracis/Bacillus cereus group (19); multiple species from the family Clostridiales such as Clostridium acetobutylicum (20), Clostridium perfringens (21), and Clostridium botulinum (22); and the important human pathogen Clostridium difficile from the family Peptostreptococcaceae (23) (see Galperin et al. [107] for details of the phylogeny). The first phylogenomic analysis of sporulation was performed by Stragier, who analyzed the distribution of 125 sporulation genes across the five endospore-forming genome sequences (B. subtilis, B. anthracis, Geobacillus stearothermophilus, C. acetobutylicum, and C. difficile) available at the time (24). One of his principal observations, corroborated by Stephenson and Hoch (25), was that, although the master regulator of sporulation, Spo0A, is present in all Clostridium spp., the phosphorelay that initiates sporulation in B. subtilis by phosphorylation of Spo0A is not conserved. This suggests that the triggers for sporulation are likely to vary considerably from species to species, especially among Clostridium spp., where at least two different pathways of direct phosphorylation of Spo0A by histidine kinases are known to exist (26) (see also Dürre [108]).

Conversely, the sporulation sigma (σ) factors appear to always be conserved (σ is the subunit of the RNA polymerase responsible for recognizing core promoter regions on the DNA). Shortly after the initiation of sporulation (Fig. 1), the asymmetric division of the sporulating cell near one pole generates two cell types:
a forespore, which will mature into a spore, and a larger mother cell, which is necessary at every stage of forespore maturation. Each cell will activate cell type-specific σ factors: σ^F and σ^G in the forespore; σ^E and σ^K in the mother cell. To a large extent, the sequential activation of the sporulation σ factors is conserved in the characteristic order: σ^F → σ^E → σ^G → σ^K (27). The spatial and temporal coordination of this regulatory cascade is achieved by several intercellular signaling mechanisms discussed by Dworkin (109). Unexpectedly, recent studies have revealed that the temporal coordination of the sporulation σ-factor cascade is less strict in Clostridium spp., in the sense that dependency on the preceding σ factor for activation of the subsequent σ factor is not absolute (28–30). In C. difficile, σ^G activity is independent of σ^E, while σ^E activity is partially independent of σ^F. Furthermore, consistent with the original observation that σ^K activity is controlled by a mother cell-specific excision of an intervening sequence element of phage origin in the sigK gene, σ^K activity does not require σ^G in C. difficile (31).
Transcriptional profiling experiments in *B. subtilis* (discussed below) vastly expanded the number of annotated sporulation genes and resulted in further phylogenetic analyses. For example, de Hoon et al. (32) traced the presence of *B. subtilis* sporulation genes in 24 spore-forming species to investigate the extent of the sporulation network conservation. The principal conclusion of this article was that, while there were differences in whether or not a gene was present in a given species, if the gene was present, its regulation was usually conserved (this was especially true of spore coat genes). Another study by Xiao et al. (33) compared the known and putative spore germination genes of 12 *Bacillus* and 24 *Clostridium* spp. and concluded that the number of operons encoding germination receptors was lower in *Clostridium* than in *Bacillus* spp. This suggests that *Clostridium* spp. have different germination strategies that may involve yet-to-be-identified Ger proteins, especially in *C. difficile*, where no germination receptor has been characterized thus far (23), even though glycine and bile salts have been identified as cogerminants (34).

While these studies have all depended on *B. subtilis* to provide the initial set of sporulation genes used for comparative genomics, more recent work has incorporated information from other spore-forming species. Additional sporulation genes were uncovered in phylogenetic profiling approaches. Specifically, by assessing the genomes of 46 sporeformers, eight genes previously unrecognized as sporulation genes in *B. subtilis* were found to be significantly enriched in spore-forming bacteria (35, 36). In addition, Galperin et al. (37) aimed to identify the minimal set of genes that are essential for sporulation by evaluating the distribution of 651 genes that are preferentially expressed during sporulation. This compendium of genes was compiled from phenotypic characterization of sporulation mutants, transcriptional profiling experiments in both *B. subtilis* and *C. acetobutylicum*, and proteomic analyses of spore content. In toto, substantial differences were noted between *Bacillaceae* and *Clostridiales* sporulation programs, particularly when it came to regulation of the onset of sporulation and the composition of the spore coat. Since numerous sporulation and germination genes characterized in *B. subtilis* do not appear in the genomes of *Clostridium* spp., sporulation and germination studies in *Clostridiales* should be expanded to better understand the differences between the two main classes of endospore-forming species.

### The *B. subtilis* Transcriptome and Regulators of Gene Expression

#### Transcriptomics of Sporulation

Gene arrays have existed since the 1970s in the form of dot blots, but it was the development of gridding robots in the 1980s and the automation of PCR in the 1990s that allowed for transcriptional profiling on a global scale (38). For *B. subtilis*, the first genome-wide array experiment was performed just a few years after the publication of the genome sequence, with macroarrays constructed by PCR amplifying all the identified open reading frames and spotting the cDNA onto a membrane (39).

The initial genome-wide macro- and microarray experiments in *B. subtilis* were designed to identify genes that were activated or repressed during the transition from vegetative growth to stationary phase (39–41). Fawcett et al. (39) compared wild-type cells with cells unable to produce Spo0A~P and found 586 genes with altered expression profiles during the early stages of sporulation in a spo0A mutant. By profiling the sigF mutant and using computational searches to discover conserved upstream sequence motifs for σ-factor binding, genes were identified as putatively under the control of σF in the forespore or σE in the mother cell. Similarly, Caldwell et al. (41) compared global gene transcription from a wild-type culture to a culture with a knockout of the gene encoding the transition state regulator ScoC. In this study, it was found that ScoC directly or indirectly affected the expression of 560 genes, in particular genes involved in motility, sporulation, competence, and degradative enzyme production. The transition to stationary phase was also profiled by comparing gene transcription in wild-type cells against cells lacking the stationary phase sigma factor σH (40). In this experiment, Britton et al. (40) found 87 σH-controlled genes and discovered that, in addition to its previously characterized role in sporulation and competence, σH was involved in the regulation of cytochrome biogenesis, transport, cell wall metabolism, and generation of potential nutrient sources.

Comprehensive transcriptional profiling of the mother cell and forespore lines of gene transcription aimed to identify regulators for each known sporulation regulator (42–44). For the mother cell, mutants were constructed to elucidate the regulons in the mother cell hierarchical regulatory cascade of \( \sigma^E \rightarrow SpoIID/\sigma^H \rightarrow GerR \rightarrow \sigma^E \rightarrow GerE \) in strain PY79 (42). This also meant inactivating the downstream regulator in the cascade to limit indirect gene expression effects. Prior to this experiment, little information was available about the full...
set of genes governed by these mother cell regulators, with the exception of σE which had been previously profiled \( (45, 46) \). It was seen that approximately 9% of the genes in the \( B. subtilis \) genome are turned on specifically in the mother cell and that the mother cell program of gene expression consists of a series of pulses where a large number of genes are activated and then shortly after turned off by the next regulatory protein in the hierarchy. Some genes, however, show protracted expression over the course of sporulation and are under the dual control of more than one σ factor, including a few cases of genes simultaneously expressed in both the mother cell and forespore \( (47, 48) \). Subsequently, the method in reference 42 was applied to the forespore regulators to understand the forespore line of gene expression: σE → RsfA → σG → SpoVT \( (43) \).

More recently, Nicolas et al. \( (49) \) collected a large transcriptome data set for strain 168 by hybridizing 269 tiling arrays over 104 diverse experimental conditions, including aerobic and anaerobic cultures, exploiting various energy and carbon sources, in media promoting growth, motility, biofilm formation, competence, or sporulation. Transcription was detected for 85% of the annotated genes under one or more conditions. In addition, 1,583 previously unannotated RNAs were identified, and it was discovered that 13% of protein-coding sequences overlapped with antisense RNAs. Importantly, the number of \( B. subtilis \) promoters was found to be three times higher than previously thought, with about half of protein-coding sequences transcribed from multiple promoters. Approximately 91% of promoters had an identifiable putative σ factor-binding motif, and σ-factor activity was shown to be responsible for 66% of the variance in transcriptional activity (i.e., the condition-dependent changes in gene expression that were observed in this data set of 104 experiments). The variance was not consistent for each σ factor; sporulation σ factors had the highest values, while the major σ factor in \( B. subtilis \), σA, had low variance. These observations suggest that σA-controlled promoters rely more heavily on other transcription regulators for condition-specific regulation.

Transcriptional profiling experiments have also been performed on many other spore-forming species. In fact, \( B. anthracis \) was the first bacterium to have a comprehensive single-nucleotide resolution view of its transcriptome using a high-throughput RNA sequencing (RNA-seq) approach \( (50) \). An extended transcriptional analysis of the \( C. acetobutylicum \) life cycle has been completed using DNA microarrays \( (51) \). In this study, 24 time points were sampled from exponential growth into sporulation, allowing for the investigation of sporulation dynamics. These data allowed for the first comparison of the temporal orchestration of the sporulation pathway between \( C. difficile \) and \( B. subtilis \) spp. More recently, genome-wide gene expression in \( C. difficile \) was analyzed by transcriptional profiling during sporulation of spo0A, sigF, sigE, spoIIID, sigG, and sigK mutants. One study used RNA-seq \( (28) \), while another study relied on DNA microarrays \( (30) \). More than 300 genes were shown to be turned on during sporulation in \( C. difficile \) \( (28) \), with >200 genes under the direct control of a sporulation σ factor \( (30) \).

### Regulators of Gene Expression in \( B. subtilis \): σ Factors, Transcription Factors, and Small RNAs

Understanding the \( B. subtilis \) transcriptome requires determining the regulons of individual regulators of gene expression. As in other organisms, \( B. subtilis \) transcriptional factors (TFs) can be identified on the basis of their structural, biochemical, and genetic properties. Structural features of TFs, most commonly DNA-binding motifs, have been deposited into Pfam, a database of protein domain families \( (52) \). With the use of the information from Pfam, the entire genome of \( B. subtilis \) was searched to identify all potential TFs, with the results of the analysis compiled into DBTBS, a database of \( B. subtilis \) promoters and transcriptional regulators \( (53) \). Currently, the repertoire of DNA-binding proteins regulating transcription in \( B. subtilis \) comprises 19 σ factors and 126 TFs. Importantly, this number only includes TFs with known target genes; therefore, the total number of transcriptional regulators in \( B. subtilis \) will increase as more TFs are characterized experimentally.

In addition to these DNA-binding proteins, non-coding RNAs regulate gene expression in \( B. subtilis \) and most likely other spore-forming bacteria \( (54) \). RNA regulators fall into three groups: riboswitches, which are part of the mRNA they regulate and carry out regulation by folding into alternative structures based on ligand availability (usually a specific metabolite); small RNAs (sRNAs), most of which bind to a target mRNA by base-pairing and modulate the translation and stability of the transcript; and CRISPRs, which have been shown to interfere with bacteriophage infection. The first sRNAs were identified in the 1980s; however, the prevalence of sRNAs and their contributions to numerous physiological responses was not realized until the early 2000s. Between 2001 and 2002, several studies that performed systematic computational searches for the conservation and presence of orphan promoter
and terminator sequences in the intergenic regions of *Escherichia coli* resulted in the identification of many new sRNAs (55). The identification of sRNAs has further increased through the use of RNA-seq and tiled arrays with full genome coverage.

Several studies have been performed with the goal of identifying transcriptionally active regions of the *B. subtilis* genome. Using tiling arrays, Rasmussen et al. (56) profiled RNA collected from exponentially growing cells in both rich (Luria Bertani, LB) and minimal (M9) media and found 84 putative sRNAs. Using an RNA-seq approach, Irnov et al. (57) collected samples in glucose minimal medium during stationary phase, further increasing the number of sRNA candidates to over 100. The first sRNAs found to be under sporulation control were also discovered by using tiling arrays. Silvaggi et al. (58) identified three sRNAs: one under indirect Spo0A~P control, another under εG control, and the final under εK control. Using a similar approach, Schmalisch et al. (59) identified two additional sporulation sRNAs. The exact role played by these sRNAs during sporulation in *B. subtilis* and their mechanism of action remain to be determined.

**GENE FUNCTION IN A SYSTEMS-WIDE CONTEXT**

**Essential Genes**

Since the beginning of the genomics era, one specific goal of systems biology has been to identify the minimal set of genes required for growth of an organism. One of the first studies to do this was performed by using the Mollicute *Mycoplasma genitalium* (60). This bacterium was chosen because it is a self-replicating organism with a small genome containing about 500 protein-coding genes. With the use of global transposon mutagenesis, it was found that about 300 genes were essential under laboratory growth conditions. Another set of essential genes was compiled for *B. subtilis* by systematically inactivating each gene and observing if the organism would grow in rich medium (61). It was found that 271 of the 4,100 protein-coding genes annotated at the time of that study were necessary for growth; however, this number was revised recently following further analyses characterizing novel essential genes and, at the same time, showing that several genes thought to be essential are actually dispensable. The new total for *B. subtilis* stands at 254 genes (62). A modernized version of the global transposon mutagenesis approach (named *Tn-seq*) was developed to accurately map essential genes by assembling a transposon library and measuring the changes in frequency of each mutant strain and determining the sites of transposon insertion by sequencing flanking regions. When combined with measurements of fitness, quantitative genetic information can be obtained in addition to gene essentiality data (63). This strategy has not yet been applied to *B. subtilis*.

**Global Phenotypic Approaches: Synthetic Genetic Arrays, Chemogenomics, Proteomics, Protein-Protein Interactions, and Protein Localization on a Large Scale**

While minimal genome studies are informative, essentiality is only one of the many criteria that can be used to characterize gene function. Based on the construction of gene deletion libraries, methods have been developed to further explore gene function by analyzing the phenotypes of double-gene knockouts (synthetic genetic arrays [SGAs]) and by investigating the fitness of single-gene deletions under a variety of chemical challenges (chemogenomics).

SGAs were initially developed in yeast. In this method, a query mutation is crossed to a deletion mutant library and the resulting double mutants are scored for fitness defects (64). With the use of this technique in the budding yeast *Saccharomyces cerevisiae*, multiple relationships between genes have been identified, including quantitative genetic interactions based on growth rate measurements (64–66).

To perform large-scale phenotypic profiling in *E. coli*, Nichols et al. (67) implemented a chemogenomics strategy. A single-gene deletion library, completed with partial-loss-of-function hypomorphs of essential genes and sRNA knockouts, was exposed to many chemical stresses (including challenges by antibiotics, detergents, dyes, and other molecules). The growth rate of each mutant strain was assessed to construct a data set including over 300 conditions and ∼4,000 genes. By analysis of phenotypic signatures, functions were predicted for previously uncharacterized genes based on the assumption that highly correlated genes are likely to function together. By revealing which cellular function (e.g., replication or cell division) is enriched in gene clusters that correlate with application of a specific chemical challenge, a possible mode of action can be suggested for a variety of drugs. As of yet, neither SGA nor chemogenomics analyses have been performed on *B. subtilis* or any other spore-forming bacterium.

Another way to probe gene function on a global scale is to study by proteomics the occurrence of gene
products in response to specific conditions. Proteomics generally refers to the analysis, by mass spectrometry, of the entire complement of proteins and protein modifications present in a sample. For example, mass spectrometry on mature *B. subtilis* spores identified the presence of 69 proteins not previously known to be present in the spore (68). Similarly, proteomic analyses of spore coat extracts revealed three new coat proteins in *B. subtilis* and six candidate spore coat proteins in *B. anthracis* (69). Nineteen additional candidates were reported in another study; however, about half of these candidates are produced from genes expressed in the *B. subtilis* forespore and are therefore unlikely to be genuine coat proteins, since coat proteins are always synthesized in the mother cell (70).

Another proteomics-based method called SILAC (stable isotope labeling with amino acids in cell culture) can provide information about relative protein concentrations and be used to compare two samples (e.g., a mutant strain and a wild type under similar conditions) (71). In this method, one sample is grown in a medium containing standard amino acids and another sample is grown with amino acids containing heavy isotopes. The samples are then pooled, trypsinized, fractionated, and analyzed with mass spectrometry. Relative protein concentrations are obtained by quantifying the difference between otherwise identical peptides that differ in mass. The SILAC strategy was applied to observe the differences in the proteome of *B. subtilis* cells grown in succinate or starved for phosphate (72); changes in levels of over 1,500 proteins were quantified in that manner. Much would be gained from expanding this work to a larger set of conditions and comparing proteomics data sets with transcriptomics data sets collected under identical conditions.

Studying the physical interactions between proteins can also aid in the elucidation of gene function, because proteins that interact with each other are likely to play a role in the same biological processes. Protein-protein interactions can be investigated on a large scale by using yeast or bacterial two-hybrid assays. In the largest interactions can be investigated on a large scale by using a spectrum and spectra are connected to similar nodes by edges.

Similar to protein-protein interaction, protein colocalization can provide supporting evidence for functional predictions. For example, protein localization was studied in budding yeast by constructing fluorescent protein fusions for 75% of the proteome (74). Similarly, a high-throughput cloning and microscopy pipeline was developed to observe both N- and C-terminal fluorescent protein fusions in *Caulobacter crescentus*, resulting in the identification of nearly 300 proteins with specific subcellular localization in this organism (75). While no global analysis of protein localization has been performed in a spore-forming bacterium, a protein localization analysis of the spore coat proteome has been done in *B. subtilis* (76, 77).

**Metabolomics (nanoDESI-MS, Annotation of Metabolic Networks)**

Another type of global phenotypic approach is metabolomics, where mass spectrometry is used both to determine the spectrum of chemical compounds produced by bacterial cultures and to analyze how this chemical signature is influenced by changing environmental conditions or by mutation of specific genes. Watrous et al. (78) aimed to study the metabolic exchanges that microbes participate in to communicate with their local environments, including neighboring microbes. To do this they developed a method called nanospray desorption electrospray ionization (nanoDESI)-mass spectrometry (MS) as a way to study the compounds secreted from bacterial colonies. This process reveals many classes of compounds in a single data set; however, because no database was available for molecules involved in metabolic exchange, a novel procedure had to be designed to simplify and explore the nanoDESI-MS output. A network-based workflow was developed in which the data were first simplified by merging identical spectra (i.e., spectra exhibiting identical mass-to-charge [m/z] ratios and fragmentation patterns). Next, the similarity for each possible pair of spectra was calculated based on the precursor m/z ratio difference and the relative intensities of the fragment ions. This information was then integrated into a network viewable in Cytoscape (an open source platform for visualizing networks, discussed below), where each node represents a spectrum and spectra are connected to similar nodes by edges.

Experiments performed with nanoDESI-MS show that the method is able to recover known molecular signals and can be used to analyze microbial communication. Using *B. subtilis* strain NCIB3610 to test the procedure, molecules known to be secreted during biofilm formation (such as surfactin) and variants of these compounds were observed to cluster together in the network. A 60-hour time series of a NCIB3610 colony was done to observe whether the metabolic
output changed over time. Increases in surfactin, plipastatin, and subtilosin production were noted, as well as a major lipid production shift. This study also analyzed the interface of colonies of different species. While the interface of *Streptomyces coelicolor* and *B. subtilis* strain PY79 colonies is characterized by a known increase in SapB in *S. coelicolor* and SKF in *B. subtilis* (79), many previously unknown signals were also detected. Overall, this method will aid in furthering our understanding of how microbes communicate with other species by providing a way to study compounds that act as mediators of cell-cell signaling. Importantly, this approach can also be used to link the production of signaling compounds to the presence of a particular gene or group of genes, and thus contribute to the discovery of gene functions.

The extensive gene function information gathered from the application of genome-scale and more traditional methods can then be used to identify orthologous genes in other organisms, where they are likely to perform similar functions. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a large database that performs gene annotation using a cross-species annotation method (80), where the function of a gene experimentally characterized in one species is putatively applied to homologous genes found in other species by searching all available genomes.

For the annotation of metabolic pathways, Plata et al. (81) have designed a novel method that considers context associations to improve upon homology-based annotations. Specifically, because genes involved in the same metabolic pathway are likely to display similar patterns of expression and conservation across species, this probabilistic approach takes into account coexpression, chromosomal location, and phylogenetic distribution during the annotation procedure. This method has better precision and recall than more classic procedures, especially in cases of low sequence homology. Several predictions were experimentally validated, including those for the *sps* (spore polysaccharide synthesis) operon, whose products are involved in rhamnose synthesis, a known component of the spore surface (82).

**MODELING REGULATORY NETWORKS**

A key objective of systems biology is to take advantage of transcriptome data sets similar to those described above to construct whole-cell dynamical models of transcriptional regulation. Such models can potentially be thought of in two ways: as a global network, where every regulator (TF or sRNA) is connected to the gene(s) it regulates, or as a system of equations that, given a starting condition, can predict transcriptome-wide responses to defined perturbations. The global network is a comprehensive map of gene regulation hypotheses that can be used to identify new targets of known regulators, while the system of equations can reveal how the expression of each gene in the model will be affected following a specific environmental or genetic modification. When possible, models should be supplemented with other data types (e.g., proteomics) to help resolve instances where regulatory mechanisms rely on postranscriptional events. In addition, chromatin-immunoprecipitation data (from ChIP-seq or ChIP-on-chip experiments that can reveal where a given TF binds in the chromosome) are especially helpful to distinguish between direct and indirect effects of specific TFs.

**Models of Central Metabolism**

Initial efforts in network construction and analysis of system-wide responses to perturbations focused on the central metabolism in both *E. coli* and *B. subtilis*. Ishii et al. (83) measured gene expression levels in various mutant strains of *E. coli* with disruptions in genes encoding enzymes involved in central carbon metabolism (i.e., the glycolysis, tricarboxylic acid cycle, and pentose phosphate pathways). The cultures were grown in a chemostat over five different dilution rates, from nearly glucose-starved to almost unlimited glucose concentrations. Comparisons of mRNA, protein, and metabolite levels showed that the metabolic network of *E. coli* is highly robust. While it was seen that an increase in growth rate resulted in a global increase in mRNA and protein levels (likely to meet the increased metabolic need), single-gene disruptions had little effect on the regulation of mRNA and protein levels. The overall stability of the network implies that rerouting of metabolic fluxes is favored over large changes in enzyme levels as a response to a genetic perturbation.

In *B. subtilis*, Goelzer et al. (84) constructed a manually curated model of the central metabolism derived from public data and expert knowledge. All the biochemical reactions in the metabolic network were included, along with all the known levels of regulation involved in metabolic pathways. The final model included 563 reactions, 534 genes, and 456 metabolites. Another model by Oh et al. (85) reconstructed a larger network, supported by growth phenotypes of knockout strains, composed of 1,020 reactions, 844 genes, and 988 metabolites. With the use of these models to
examine the organization of metabolic regulation, a clear hierarchy emerges, where local and global regulators can be readily distinguished. Typically, local regulators control one specific pathway and have potential metabolite effectors that belong to the controlled pathway (either as the end product or as an intermediate metabolite of the pathway). For instance, the tryptophan synthesis pathway is dependent on several local regulators and effectors, including tryptophan itself (86). Global regulators, such as CcpA (the main regulator for carbon catabolite repression), TnrA (the main regulator for nitrogen metabolism), and CodY (a regulator of many pathways including branched amino acid synthesis), are defined as nonlocal regulators in the sense that they are involved in the regulation of many pathways. Thus, in response to environmental changes, global regulators ensure coordination within the system, while local regulators provide a second level of regulation.

Inferring Regulatory Relationships from Large-Scale Gene Expression Data Sets
In 2007, Bonneau et al. (87) constructed a predictive model of the global transcriptional regulatory network in the archon Halobacterium salinarum. This was achieved through the integration of transcriptomics and proteomics data sets, collected under a variety of genetic and environmental perturbations. New TFs were identified with protein structure predictions, whereas new operons and coregulated groups were inferred by computational analysis of genome structure. The model covered approximately 80% of the H. salinarum transcriptome and accurately predicted many novel transcriptional regulatory relationships.

To implement an integrative biclustering strategy, the algorithm cMonkey was developed. Biclustering is the grouping of genes coregulated under subsets of experimental conditions, whereas integrative biclustering incorporates additional data sources such as conserved upstream sequence motifs, genomic position, and previous information gathered from manually curated databases, e.g., KEGG. As a result, coregulated—and by extension cofunctional—groups of genes (biclusters) can be obtained via integration of expression data, cis-regulatory motifs, evolutionarily conserved associations among proteins, and metabolic pathways. Next, the algorithm Interelator was used to predict which regulators affect the expression of these biclusters, as well as regulatory relationships between TFs and genes not included in biclusters. By the use of the model as a driver for hypothesis generation and subsequent testing, experimental validations have confirmed a number of these regulatory hypotheses, including a hierarchy of regulation for copper efflux, and assigned functions to proteins with no characterized orthologs (87).

To learn networks of conserved coregulated gene groups using multiple-species data sets, Waltman et al. (88) present an algorithm for multispecies integrative biclustering based on the cMonkey biclustering algorithm. Multispecies cMonkey (MSCM) adds another layer to traditional biclustering approaches through the consideration of evolutionary conservation, the concept being that modules composed of conserved genes are more likely to be biologically significant. MSCM biclustering was performed on all the orthologous gene pairings between B. subtilis, B. anthracis, and Listeria monocytogenes. Analysis of the biclusters led to the unexpected discovery that B. anthracis, although nonflagellated and nonmotile, expressed many flagellar motility-chemotaxis genes. Further investigation revealed that the loss of motility was likely due to the accumulation of frameshift mutations in a few of the key genes required for flagellar assembly. Another interesting result was the species-specific timing of expression of a subset of sporulation genes. Genes expressed early in the mother cell in B. subtilis (under the control of the sporulation sigma factor $\sigma^E$) were divided into three biclusters. The splitting of the $\sigma^E$-dependent genes into three biclusters was due in part to functional categorization, but primarily to the three biclusters having different expression profiles in B. anthracis. In particular, one of these biclusters, composed of genes with related metabolic functions, is expressed earlier than the other two biclusters, and even before expression of $\text{sigE}$. This observation suggests a partial genetic rewiring in the sporulation program, implying that transcription of this bicluster in B. anthracis is controlled by regulators other than $\text{sigE}$.

Steps Toward a Global Gene Regulatory Network Model for B. subtilis
To organize B. subtilis genes into putatively coregulated modules, Fadda et al. (89) used the integrative biclustering algorithm DISTILLER. Coexpression data from 213 publicly available microarray data sets from five studies, performed on 10 platforms and covering six different experimental conditions (DNA damage, heat challenge, peroxide stress, phosphate starvation, quorum sensing, and sporulation), were combined with sequence motif information to obtain 142 modules containing 1,153 of the genes in the B. subtilis genome (approximately a quarter of the protein-coding genes).
The regulators for each module were first determined based on the presence of upstream motifs known to be recognized by specific regulators. When this option was unavailable, assignment of a regulator to a specific module relied on the observation that many regulators autoregulate, or regulate adjacent genes. Therefore, putative regulators were assigned based on the genomic distance between a gene encoding a regulator and a gene containing the motif of interest. This also means that any regulator present in a module would automatically be considered a putative regulator of that module. In this model, 44% of the known regulatory interactions were recovered and 417 new targets were predicted. Although several new regulators were proposed, no experimental validation was performed to follow up on those predictions.

To search for regulatory relationships between σ factors and groups of cotranscribed genes in B. subtilis, Nicolas et al. (49) estimated pairwise correlations between promoter activities and transcriptionally active regions (TARs) in their transcriptomics data set (see above). TARs were distributed in coregulated modules by hierarchical clustering of the transcriptional profiles. To systematically identify σ-factor regulons, an unsupervised algorithm was developed to combine DNA sequence information upstream of the transcribed region (i.e., the putative promoter region) with the clustering of transcriptional profiles to determine the conserved bipartite motif recognized by each σ factor. By this method, ~91% of the TARs had an identifiable putative σ-factor-binding site in their upstream regulatory region. In order to determine the participation of σ factors in transcriptome plasticity, the average transcription signal for all the promoters of a cluster was estimated and a linear relationship was assembled between the cluster activity and individual promoters. It was found that 66% of the changes in gene expression over the set of conditions present in the data set were due to variations in σ-factor activity. The role of other TFs (that may act in combination with the major sigma factor, σA) was not touched upon.

A project is also currently underway to infer the global transcriptional regulatory network of B. subtilis (A. R. Bate, M. Arrieta-Ortiz, C. Hafemeister, T. Chu, A. Greenfield, S. N. Barry, R. Bonneau, and P. Eichenberger, unpublished data). This project was initially inspired by the strategy discussed above to obtain the H. salinarum network (87). By using two large B. subtilis gene transcription data sets (one collected specifically for the project, the other from Nicolas et al. [49] and Buescher et al. [90]) and an updated version of the Inferelator that considers known network edges and transcription factor activities to aid in making predictions (91), a global transcriptional regulatory network will be constructed with all known transcriptional regulators (σ factors and TFs). This new model will also include putative regulators, such as proteins with predicted DNA-binding domains and sRNAs. Upon completion, the network will be made available to the scientific community to be used as a resource for hypothesis generation. A key difference between this work and the work on H. salinarum is that it takes advantage of >2,500 previously known regulatory relationships archived in SubtiWiki (see below) (92, 93). Incorporation of this prior information will dramatically improve the accuracy of the resulting network as previously demonstrated (91). The present project may also more efficiently reveal the inherent biases of the network inference approach by providing a precise measurement of the error rates (94).

**Network Visualization**

The multiplication of -omics data sets has necessitated the development of tools to store, process, and visualize the information available. The SubtiList database was assembled as a companion to the B. subtilis genome sequencing project discussed above (5, 95). This database includes the annotated genomic sequence and several sequence analysis and visualization tools. Newly found annotations and functional assignments were added in subsequent versions, but the last update was made in 2002 (96). SubtiList was then integrated into GenoList, a database dedicated to querying and analyzing genome data from many bacterial species (97). Recently, a new wiki-based resource called SubtiWiki has been created to store comprehensive, up-to-date information on B. subtilis genes, proteins, regulatory interactions, protein-protein interactions, pathways, and gene transcription profiles (92, 93). Because the site is a wiki, qualified scientists can contribute to the resource, allowing the site to stay up to date and allowing the collective knowledge of the B. subtilis community to be made accessible to everyone.

Cytoscape is a software package used for network visualization, data integration, and analysis, designed with systems biology research in mind (98, 99). Data can be displayed with the visual aspects of the network decided by the user. For example, a regulatory network can be generated, where genes are circular nodes and transcription factors are triangular nodes. Directional edges would connect transcription factor nodes and gene nodes based on gene expression data from microarray...
experiments, with the color of the edge indicating activation or repression and the width of the line reflecting the confidence score of the regulatory interaction. Gene or protein networks can also be designed with edges showing that the genes are in an operon, or that the proteins are known to physically interact. Importantly, Cytoscape affords connections to large numbers of tools (called plugins) for annotation, clustering, and visualizing networks (e.g., one plugin, described below, allows users to connect the network view presented by Cytoscape to several other tools using the Gaggle framework).

As systems biology data sets have increased in complexity, it has become necessary to use more than one visualization tool to include different types of data. Gaggle is a Java program that broadcasts gene, network, and data selection between various tools, allowing users to coordinate visualization of different aspects of gene groups, networks, and data (100). The various tools the Gaggle communicates with are referred to as geese; geese made compatible with the Gaggle include Cytoscape, Cytoscape plugins, a Global Synonym/Ortholog Translator, and several other tools (such as a data matrix and annotation viewers) that facilitate the exploration of -omics data. Using the Gaggle, a researcher is able to take a list of genes and retrieve the gene annotations, expression in a set of experiments, and a Cytoscape network showing known associations between the genes.

We previously developed the Comparative Microbial Modules Resource (CMMR) to integrate the above tools into a single interface to the analysis, data, and metadata for B. subtilis genomics. This resource is presented to the user as a website specific for the analysis of multispecies biclusters (101). It provides a way to search for multispecies biclusters that contain a gene of interest or are enriched for a specific annotation. Bicluster information includes the list of genes in that bicluster and their annotations, the corresponding gene expression profiles, enriched sequence motifs, and a functional enrichment summary. Furthermore, the resource utilizes the Gaggle to explore biclusters of interest and facilitate navigation between tools.

**Combination of Large-Scale Approaches to Obtain Whole-Cell Models**

A study from the BaSysBio consortium combined the use of multiple -omics platforms to observe changes during metabolic shifts from glucose to malate and vice versa (90). Both short-term and long-term time series data were collected for transcripts, protein levels, metabolite abundance, and promoter activity. The integration of these data types led to a detailed understanding of the metabolic shift beyond the sole measurement of transcript levels. This study was also able to uncover new transcriptionally active regions (including sRNAs), provide putative functional assignments for genes of previously unknown function, and identify posttranscriptional regulatory events to achieve a more comprehensive description of central metabolism in B. subtilis.

By observing 300 genes for which transcript, protein, and promoter activity data were available, 110 posttranscriptional regulation events were identified. More evidence for posttranscriptional events was gathered by performing a network component analysis for 154 transcription factors and their 1,754 known targets and observing the deviations between predicted transcription factor activity and measured mRNA abundance. It was inferred that 51 transcription factors had activity that was regulated posttranscriptionally; however, for 39 of these 51 transcription factors, the effector of posttranscriptional regulation remains unknown. Finally, it was seen that, while shifting to glucose metabolism involved more transcriptional regulation, posttranscriptional regulation played an important role in shifting to malate. This was determined by correlating the time course of the metabolic flux with corresponding enzyme abundance; a positive correlation was indicative of transcriptional regulation influencing the reaction rate. Modeling substrate usage and enzyme maintenance indicated that a delay in glucose uptake (shown to be dependent on two operons) could provide an evolutionary advantage when malate was the primary substrate. This last observation makes sense considering that plant roots, which constitute a preferred habitat for B. subtilis (102), often secrete carboxylic acids such as malate (see also Mandic-Mulec et al. [107]).

Karr et al. (103) present the first whole-cell computational model of an organism using transcriptome, proteome, and metabolome data from M. genitalium. To construct the model, the authors first divided the cell into 28 functional modules (including replication, transcription, translation, metabolism, and cytokinesis). Each module was then modeled using the mathematical representation deemed most appropriate, and was built, parameterized, and tested independently. The resulting submodels were integrated into a complete whole-cell model by performing simulations that considered time and cell variables (such as DNA, RNA, protein, and metabolite levels and other variables including geometry and mass). The simulations were executed by running through a loop, the start of which was the beginning of
the cell cycle. In these simulations, the submodels ran autonomously at each time step, but dependently on the cell variables determined by the various submodels at earlier time steps. If the simulation did not end in cell division, the loop began again and the simulation that led to full cell division was the one that resulted in the integration of the submodels into the full cell model. The model was validated by verifying that it recovered known information about *M. genitalium* and also by comparing it against many independent data sets not used in the construction of the model. It was found that the model indeed recapitulated many aspects of both the training data and independent experimental data across multiple biological functions.

An important feature of any model is not only to reiterate known information, but also to provide novel insights that can be experimentally tested. The *M. genitalium* model was able to predict features related to chromosomal occupancy by DNA-binding proteins and protein-protein collisions on the chromosome. The model also predicted that progress of DNA replication is controlled in part by the content of dNTPs at the start of replication, and therefore that availability of dNTPs is an emergent control of cell cycle duration, independent of genetic regulation. Simulations were performed in which each of the 525 genes in the *M. genitalium* genome was disrupted one by one, and 284 genes were predicted to be essential. Thus, the model agrees with the observed gene essentiality at 79% accuracy (see above); cases in which the model disagrees with the experimental results were explored further.

**CONCLUSIONS**

In this article, we have reviewed the impact that -omics techniques have had on the study of *B. subtilis* and other bacteria and have introduced new methods that can be used to expand the study of spore-forming organisms. While genomics and transcriptomics have been widely used to examine sporeformers, other global measurements to elucidate gene functions (such as proteomics, metabolomics, synthetic gene arrays, and chemical genomics) have been less utilized. Because these methods are shown to expand our knowledge of other microbes, it is likely that they will soon be applied to the study of *B. subtilis* and other spore-forming bacteria.

**ACKNOWLEDGMENT**

We thank Jonathan Dworkin for critical reading of the manuscript.

Work in P.E.’s laboratory is supported by grant GM081571 from the U.S. National Institutes of Health.

**REFERENCES**


Biosyst
93. Mader U, Schmeisky AG, Florez LA, Stulke J.
92. Flórez LA, Roppel SF, Schmeisky AG, Lammers CR, Stülke J.