Journal of Bacteriology

Structure and Complexity of a Bacterial Transcriptome

Karla D. Passalacqua, Anjana Varadarajan, Brian D. Ondov, David T. Okou, Michael E. Zwick and Nicholas H. Bergman

J. Bacteriol. **2009, 191(10):3203. DOI: 10.1128/JB.00122-09. Published Ahead of Print 20 March 2009.**

JOURNAL OF BACTERIOLOGY, May 2009, p. 3203–3211 Vol. 191, No. 10
0021-9193/09/$08.00+0 doi:10.1128/JB.00122-09

Structure and Complexity of a Bacterial Transcriptome▽†

Karla D. Passalacqua,[1] Anjana Varadarajan,[1] Brian D. Ondov,[1] David T. Okou,[2] Michael E. Zwick,[2] and Nicholas H. Bergman[1,3]*

School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332[1]; Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia 30322[2]; and Electro-Optical Systems Laboratory, Georgia Tech Research Institute, Atlanta, Georgia 30332[3]

Received 29 January 2009/Accepted 6 March 2009

Although gene expression has been studied in bacteria for decades, many aspects of the bacterial transcriptome remain poorly understood. Transcript structure, operon linkages, and information on absolute abundance all provide valuable insights into gene function and regulation, but none has ever been determined on a genome-wide scale for any bacterium. Indeed, these aspects of the prokaryotic transcriptome have been explored on a large scale in only a few instances, and consequently little is known about the absolute composition of the mRNA population within a bacterial cell. Here we report the use of a high-throughput sequencing-based approach in assembling the first comprehensive, single-nucleotide resolution view of a bacterial transcriptome. We sampled the *Bacillus anthracis* transcriptome under a variety of growth conditions and showed that the data provide an accurate and high-resolution map of transcript start sites and operon structure throughout the genome. Further, the sequence data identified previously nonannotated regions with significant transcriptional activity and enhanced the accuracy of existing genome annotations. Finally, our data provide estimates of absolute transcript abundance and suggest that there is significant transcriptional heterogeneity within a clonal, synchronized bacterial population. Overall, our results offer an unprecedented view of gene expression and regulation in a bacterial cell.

Although more than a thousand bacterial genomes have been sequenced, our understanding of bacterial transcriptomes has lagged far behind (see the NIH database at http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=2&type=0&name=Complete%20Bacteria). The physical structure of prokaryotic transcriptomes—operon linkages and transcript boundaries, for instance—is not defined on a genome-wide level for any species. Even though this information represents a critical step in understanding the functional and regulatory architecture of the genome, it has been explored on a large scale (>10% of the transcriptome) in only a few instances (6, 16). Similarly, although global gene expression in bacteria is routinely studied in a relative sense, where expression patterns occurring in two or more conditions are compared, there have been few attempts to comprehensively profile a bacterial transcriptome from an unbiased perspective. Consequently, little is known about the absolute composition of the mRNA population within a bacterial cell or about how individual cells' mRNA content might differ.

Recent advances in high-throughput DNA sequencing have made it possible to define nucleic acid populations at an unprecedented depth and resolution. Here we report the use of a sequencing-based approach (RNA-Seq) (17, 24) in assembling the first comprehensive, single-nucleotide resolution view of a bacterial transcriptome.

* Corresponding author. Mailing address: School of Biology, Georgia Institute of Technology, 310 Ferst Dr., Rm. 231, Atlanta, GA 30332-0230. Phone: (404) 894-8418. Fax: (404) 894-0519. E-mail: nickbergman@gatech.edu.

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

▽ Published ahead of print on 20 March 2009.

MATERIALS AND METHODS

Growth of *Bacillus anthracis* Sterne (34F$_2$). *B. anthracis* Sterne (34F$_2$) was grown in modified G medium (MGM) at 37°C (with shaking at 250 rpm) and in MGM plus 0.8% sodium bicarbonate in 14 to 15% CO_2 at 37°C (with shaking at 150 rpm). Cells were harvested throughout the bacterial life cycle, as indicated in Fig. 1A and Table 1 (our primary aim here was to maximize the number of transcripts represented within the samples that were collected). RNA was collected in four biological replicates from each of the eight points and under the conditions indicated in Fig. 1A and Table 1.

RNA isolation. RNA collection was done by hot-phenol extraction as described previously (1) (the full protocol is available at http://bergmanlab.biology.gatech.edu). All samples were processed using the Qiagen RNeasy RNA cleanup protocol with on-column DNase digestion for removal of genomic DNA. Multiple 10-μg quantities of each RNA sample were depleted of rRNA, using an Ambion MicrobExpress kit per the manufacturer's instructions. Since rRNA depletion removed most of the RNA in a given sample (rRNA constituted ≥75% of the total RNA) and the resulting mRNA yields were consequently quite low, isolates from equivalent culture conditions were pooled in equal measure to bring the total RNA sample back up to 10 μg prior to cDNA synthesis. Samples were assayed for RNA integrity on a Bio-Rad Experion automated electrophoresis station with prokaryotic RNA StdSens chips and were quantitated on a NanoDrop 1000 spectrophotometer.

cDNA synthesis. After rRNA depletion, ~10 μg of rRNA-depleted mRNA was used to make cDNA, using an Invitrogen SuperScript II double-stranded cDNA synthesis kit with random hexamers according to standard protocols. All preparations were quantitated on a NanoDrop 1000 spectrophotometer. The full cDNA preparation protocol is available at http://bergmanlab.biology.gatech.edu.

DNA sequencing. SOLiD sequencing was performed at Agencourt Bioscience (Beverly, MA) and SeqWright (Houston, TX). Library preparations, fragment library protocols, and all SOLiD-run parameters followed standard Applied Biosystems protocols. Illumina/Solexa sequencing followed standard Illumina/Solexa methods and protocols. Both systems were used because in this study our aim was simply to collect as much sequence coverage as possible, and in pursuing this goal, we utilized all available sequencing systems. We note, however, that the relative performances (overall coverage, data output, etc.) noted in Table 1 for the Illumina and Applied Biosystems sequencing platforms cannot be compared directly. This study was not intended or designed to be a direct comparison of the two systems, and interpreting the differences shown in Table 1 is difficult given the many differences in cDNA/library preparation and the rapid evolution of each system over the past year.

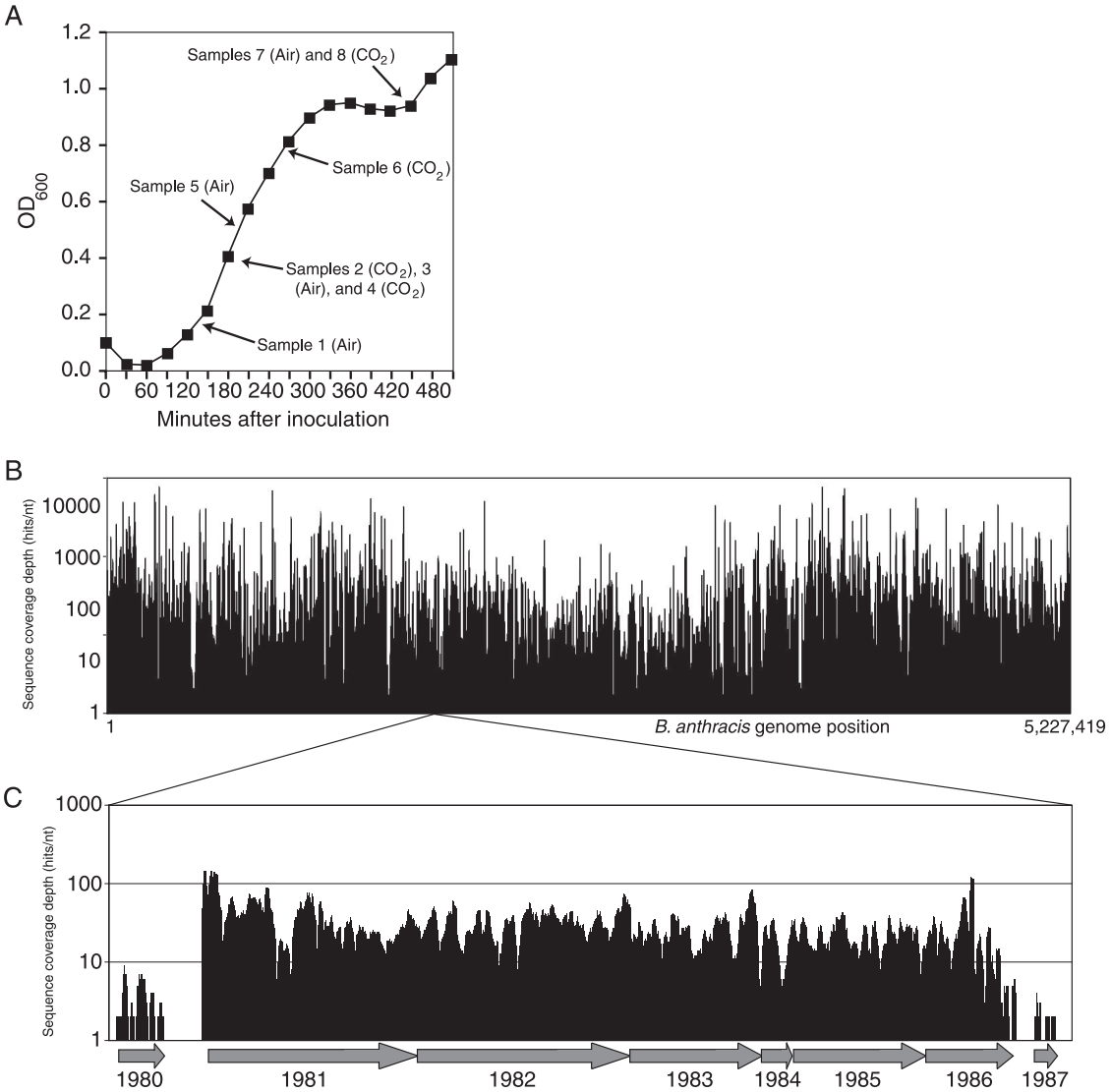

FIG. 1. Sample collection for and representative global structure of the *B. anthracis* transcriptome. (A) Representative growth curve of *B. anthracis* in MGM, with approximate RNA collection points shown by arrows. The atmosphere under which each sample was collected is indicated; note that growth rates in air and 15% CO_2 environments were similar, with slightly slower growth in CO_2. OD_{600}, optical density at 600 nm. (B) Sequence coverage across the entire *B. anthracis* genome (5.2 Mb). Data shown are from sample 5. (C) Magnified portion of the plot from panel B, showing sequence coverage over an ~10-kb region of the *B. anthracis* chromosome, with genes GBAA1980-7 indicated below by arrows.

Mapping and processing sequence data. Illumina sequence data were mapped to the *B. anthracis* Ames Ancestor genome, using SOAP (13) with 3′-end trimming and a tolerance of ≤4 nucleotide mismatches. ABI SOLiD data were mapped, using SOCS (18) with a tolerance of ≤5 mismatches. SOCS generates both a best match for each read and a genome-wide census of sequence coverage; since SOAP provides only the former, ad hoc Perl scripts (available at http://bergmanlab.biology.gatech.edu) were used to generate coverage data, using the best-match information. Once in coverage depth form, the SOLiD and Illumina data were processed identically.

Quality control and statistical and bioinformatic analysis of sequence data. Coverage data sets (including only unambiguously mapped reads) for each sample were checked to ensure that both strands' data (considered separately by SOAP and SOCS) were closely related. In all cases, we noted a Pearson correlation of >0.93; therefore, the plus- and minus-strand data were merged. Technical replicates all showed correlations of >0.99, and data were merged to yield 8 sample data sets, which were analyzed for several potential biases. Coverage depth (measured as hits per nucleotide) was compared to gene length by calculating a Spearman rank correlation coefficient between the two across the genome, with no significant correlation found in any sample. Similarly, neither the signal itself nor the local (i.e., within-gene) variance within the signal showed a significant correlation with GC content (using window sizes of 1 or 35 nucleotides [nt] for calculating GC content). Finally, we compared the average coverage depth for the 5′ quarter of each gene to the average coverage depth within the corresponding 3′ quarter, using a Mann-Whitney test, and found no significant bias.

Statistical analyses were done using R, Microsoft Excel 2008, and GraphPad Prism 5.0a. Processing of sequence data for mapping or quantification was done using custom Perl scripts that are available for download at http://bergmanlab.biology.gatech.edu.

Transcript start site identification. Transcript start sites were flagged using the following set of rules. (i) Genes with an average coverage score (i.e., an average sequence data coverage depth) of ≤0.5 were flagged as "insufficient data." (ii) Genes with an average coverage score of >0.5 and continuous coverage that extends into a codirectional upstream gene were flagged as downstream members of an operon. (iii) For all other genes, we began at the open reading frame midpoint and moved through the coverage data toward the 5′ end of the gene, as follows. If a coverage depth score of 0 was encountered, we checked the number of adjacent 0 scores. If the number was <35, we ignored it; if it was ≥35,

TABLE 1. Summary of RNA sequence data

Sample no.	OD_{600} (growth phase)[c]	Atmosphere	Total no. of sequence reads	No. of reads mapped[d]	No. of reads unambiguously mapped	Total no. of bases unambiguously mapped	Percentage of genome represented	Sequencing platform[e]
1	0.2 (early log)	Air	34,099,100	17,106,103	9,444,198	330,456,930	78.41	SOLiD
2[a]	0.4 (mid-log)	15% CO_2	28,664,981	19,760,450	3,006,469	103,392,415	61.81	SOLiD
3[b]	0.4 (mid-log)	Air	15,243,969	11,621,654	2,063,882	66,044,224	39.72	GA I
4[b]	0.4 (mid-log)	15% CO_2	15,467,681	11,847,455	1,311,078	41,954,496	37.37	GA I
5	0.5 (mid-log)	Air	34,513,145	18,063,406	12,267,396	429,358,860	79.16	SOLiD
6[a]	0.8 (late log)	15% CO_2	52,412,661	35,483,599	7,972,458	279,036,030	76.78	SOLiD
7[a]	1.0 (late sporulation)	Air	45,061,700	15,443,108	991,054	34,686,890	51.68	SOLiD
8[a]	1.0 (late sporulation)	15% CO_2	39,337,991	27,046,865	1,966,313	68,820,955	60.26	SOLiD
Total			264,801,228	156,372,640	39,022,848	1,353,840,800	93.89	

[a] Two technical replicates; data in this row are summed totals for all replicates.
[b] Three technical replicates; data in this row are summed totals for all replicates.
[c] See Fig. 1A for an example of a *B. anthracis* growth curve. OD_{600}, optical density at 600 nm.
[d] Mapped at a tolerance of ≤4 mismatches for Illumina data and ≤3 mismatches for SOLiD data.
[e] GA I, Genome Analyzer.

the gene was flagged as "insufficient data." Once outside the gene, we called the first position with a score of 0 the transcriptional start site.

Identification of "dropped" and new loci and genes with possibly incorrect start codons. Dropped and new loci were identified by using the following rules. (i) The margin used was defined as 25% of the distance between each pair of genes, or 250 nt, whichever was smaller. (ii) For each intergenic region in the *B. anthracis* genome, the candidate region was defined as the nucleotides beginning [margin] nt downstream of the first gene (i.e., the gene with a smaller GBAA number) and ending [margin] nt upstream of the next gene. (iii) For each candidate region, nonzero scores that were >100 nt from any known noncoding RNA gene were counted. If >100 nonzero scores were found and their average coverage depth was above the depth for all genes in the *B. anthracis* genome (each sample considered separately) and if the GBAA number corresponded to a gap in the known annotation (i.e., the candidate region occurred between two genes whose GBAA numbers were two apart), the region was tagged as a "dropped" locus. If the region occurred between two genes with adjacent GBAA numbers, the region was tagged as "new."

Genes that may have incorrectly been called start codons were identified by looking for loci that have significant coverage depth (greater than or equal to the median coverage depth for all genes in the genome) across the gene but very little coverage depth around the annotated start codon (>15 of the 20 positions between −10 and +10 of the gene's translational start site having a score of 0).

SYBR green qRT-PCR. For quantitative reverse transcription-PCR (qRT-PCR) validation experiments, 18 genes were chosen to represent a range of SOLiD scores in coverage depth (scores of >6,000.00 to ~20.00). Experiments were performed using Applied Biosystems Power SYBR green RNA-to-CT one-step mix (3 experimental replicates; controls included two reactions with no reverse transcriptase and one reaction with no RNA). Reactions were run on an ABI Prism 7000. Threshold cycle values were averaged.

End-point RT-PCR for operon pair assessment. End-point RT-PCR was performed using an Invitrogen SuperScript III one-step RT-PCR system with Platinum *Taq* DNA polymerase. For each pair, four primers designated A, B, C, and D were designed, whereby A and B would amplify a region within gene 1, C and D would amplify a region within gene 2, and A and D would amplify across the intergenic region if a contiguous transcript existed (see Fig. S4 in the supplemental material). Reactions were visualized on 2% agarose gels stained with ethidium bromide.

5′-End validation using template switching with extension (TSx). The 5′ (untranslated regions [UTRs]) of selected mRNA transcripts were assessed using TSx (a modified 5′ rapid amplification of cDNA ends [RACE] protocol [15]) that uses a reverse transcriptase template-switching extension step with a "step-out" PCR amplification followed by Sanger sequencing (Agencourt Bioscience Corporation, Beverly, MA). Briefly, TSx was done as follows. Invitrogen SuperScript II reverse transcriptase was used with random hexamers and a non-gene-specific 3-riboguanosine primer (TSx primer; Dharmacon RNAi Technologies) for cDNA synthesis with a TSx. The SuperScript II reverse transcriptase enzyme adds three nontemplated C's at the end of reverse transcription, and the TSx primer with three riboguanosines then adds an extended known sequence to the 5′ end of the first-strand template-switching cDNA. Three subsequent PCR amplifications were performed as follows. The first PCR (PCR1) used the TSx cDNA pool as templates, a gene-specific primer located far upstream of the ATG start site (primer TS far), and a step-out heel carrier primer (SO heel carrier) complementary to the TSx primer sequence with an additional known sequence at the 5′ end. Positive-control reactions using gene-specific primers that are located within the open reading frame sequence were used in PCR1 to test if the transcript of interest was present in the TSx cDNA pool. If the transcript was present, then PCR2 was undertaken, using a gene-specific primer (GSP) complementary to a more upstream sequence within the open reading frame (GSP inner) and a heel-Udist primer complementary to the new known sequence that was added by the SO heel carrier primer with another new 5′-end known sequence. A negative-control experiment, using only the heel-Udist primer (zero GSP primer) was also performed in order to distinguish background amplification from gene-specific amplification. These samples were then run on a 2% agarose gel and gel purified. The gel-purified DNA was used as a template for PCR3 (amplification step), using the GSP inner primer in excess of the heel-Udist primer to amplify enough DNA for Sanger sequencing. PCR products were sent to Agencourt Bioscience for sequencing using a gene-specific primer to amplify from within the open reading frame out toward the 5′ end. Raw sequence data were then analyzed by hand. A TSx UTR was defined as the sequence upstream of the ATG start site of the open reading frame in question up through the TSx and SO heel carrier primer sequences, using the *B. anthracis* Ames Ancestor genome as the reference.

The full protocol with primer sequences and touchdown PCR cycling parameters can be found at http://bergmanlab.biology.gatech.edu. Note that the TSx, SO heel carrier, and heel-Udist primers were designed with the *B. anthracis* genome in mind (i.e., they have very low homology in that genome), and so these primer sequences may not be appropriate for other bacteria.

In order to verify that the TSx protocol was able to produce valid 5′-terminus sequences, we compared the TSx-derived 5′ sequence obtained from the GBAA1981-6 operon with the transcriptional start site that had been previously determined by conventional primer extension analysis (3). The 5′ start sites matched perfectly. This locus was used in subsequent TSx versus RNA-Seq start site comparisons, and the remaining loci used for comparing start site measurements were chosen based on the following two criteria: (i) having an average sequence coverage score of >500 hits/nt and (ii) having a putative UTR based on sequencing data of >40 nt (see Fig. S3B through K in the supplemental material). All genes assayed were taken from sample 6 (Table 1).

Primer sequences. All primer sequences are available upon request.

Measurement of gene expression using sequence coverage data. The expression level for a given gene in a given sample was measured as the mean coverage depth for all nucleotides in that gene, with genes that contained repeats that did not allow for unambiguous mapping set aside and not considered further (there were <100 of these genes). For comparative purposes, coverage profiles were normalized based on the total number of unambiguously mapped reads across the genome for each sample. Microarray data used for comparative purposes were those collected by Bergman et al. (1) and archived in ArrayExpress (accession number E-MEXP-788); specific data sets used are noted in the text.

Sequence data accession numbers. SOLID and Genome Analyzer sequence data are available as both raw short read data and processed coverage plots on

the authors' website (http://bergmanlab.biology.gatech.edu) and in the NCBI GEO database (accession no. GSE13543).

RESULTS AND DISCUSSION

Definition of a bacterial transcriptome by ultra-high-throughput sequencing. Our model system for this study was *B. anthracis*, a spore-forming bacterium and the causative agent of anthrax. Studies over the past several years have defined global gene expression patterns throughout the *B. anthracis* life cycle (1), and we used those data to identify a set of eight growth conditions in which we expected that the collective transcript diversity would be maximized. Total RNA was isolated from cells harvested throughout an entire life cycle in two growth environments (Fig. 1A). After we enriched for mRNA by depleting the 16S and 23S ribosomal RNAs from our samples (see Fig. S1 in the supplemental material), RNA was converted to cDNA and subjected to shotgun sequencing, using the Illumina genome analyzer (2 samples) and Applied Biosystems SOLiD (6 samples) sequencing platforms (Table 1).

The short (~35-nt) sequence reads produced by these systems were mapped to the *B. anthracis* genome, using software tools specific to each platform (Table 1) (13, 18), with ambiguously mapped reads (i.e., those with more than one potential match in the genome) recorded separately and excluded from subsequent analyses. The unambiguously mapped reads (39,022,848, for a total of 1,353 Mbp of sequence data) were used to compile a coverage profile for each sample which reflects the depth of sequence data at each position in the *B. anthracis* genome (Fig. 1B; see also Fig. S2 in the supplemental material). As expected, technical replicates showed a very high level of correlation ($r > 0.99$), while data from separate samples showed significant differences (Spearman correlations ranged from 0.237 to 0.797) reflective of the diverse growth conditions sampled.

Collectively, we observed that roughly 94% of the *B. anthracis* genome was transcribed in one or more growth conditions, though the fraction represented by transcript sequence data in any single sample was generally much less (Table 1). As in other studies using RNA-Seq, we noticed a relatively high level of signal variance within the sequence coverage, but we did not detect any significant biases relating coverage depth to gene length, position within a transcript, or GC content (see Materials and Methods). The only bias we noted was a slightly lower average coverage depth in the several hundred kilobases directly opposite the origin of replication (Fig. 1B; see also Fig. S2 in the supplemental material). However, given that highly expressed genes in bacterial genomes tend to be found near the origin (22), this may reflect a real biological trend rather than a technical bias.

Structure of the *B. anthracis* transcriptome. As we viewed the sequence coverage for each RNA sample, the overall structure of the *B. anthracis* transcriptome appeared plainly visible, both as continuous stretches of transcription through intergenic regions putatively representing multigene operons (i.e., two or more genes transcribed on one contiguous mRNA molecule) and as distinct transcript boundaries, where coverage showed sharp transitions (Fig. 1C). These two features (operons and boundaries) have not been experimentally defined on a genome-wide scale for any bacterial species, and they are the foundation for a detailed description of a genome's regulatory architecture. Therefore, we used the sequence data to identify transcript boundaries across the entire genome. Note that although both 5′- and 3′-transcript boundaries are normally quite evident in the coverage data (Fig. 1C), we focused primarily on 5′ boundaries because regulatory mechanisms for transcription and translation in bacteria are typically mediated by sequences within or near 5′ UTRs. Hence, we examined coverage data from each sample within and around each gene, using a straightforward set of rules to identify cooperonic genes and transcriptional start sites (TSSs). In our approach, we traced sequence coverage signals upstream from the midpoint of each gene. If coverage was continuous through the upstream intergenic sequence (IGS) and into the next codirectional gene, the gene was designated as being a downstream member of an operon (2,370 genes; 41%). Alternatively, if signal coverage dropped off in the IGS upstream of the open reading frame, we designated the point at which it fell to zero as a putative TSS (3,105 identified; 54%) (see Tables S1 and S2 in the supplemental material).

As with other bacteria (16), the 5′ UTRs identified in this study were generally quite small. Of the 1,330 TSSs that were estimated with the highest confidence (i.e., TSS located in all 8 samples within a ≤40-nt window; see Tables S1 and S2 in the supplemental material), 1,164 were ≤40 nt, and only 37 were ≥100 nt (the genes with 5′ UTRs ≥100 nt are listed in Table S3 in the supplemental material). To validate our mapping approach, specifically for longer UTRs that may have regulatory functions, we used TSx (a modified 5′ RACE [see Materials and Methods]) (15) to independently determine the TSS for a select group of genes. In general, sequence coverage- and TSx-derived start site estimates matched quite closely (see Fig. S3 and Table S4 in the supplemental material). For instance, for the GBAA1981-6 (*asb*) operon UTR (Fig. 1C), the two methods indicated start sites that were 9 nt apart (Fig. 2A; see also Fig. S3A in the supplemental material). Overall, in defining TSS using both methods, 8 of 11 were within 15 nt of each other, with 5 cases having a separation of fewer than 5 nt (see Fig. S3 and Table S4 in the supplemental material).

Along similar lines, we sought to validate our approach to identifying operon structure by performing end-point RT-PCR on select gene pairs (see Fig. S4 and Table S5 in the supplemental material). With the exception of one pair that was chosen because it is highly likely to be cooperonic, the chosen gene pairs had large intergenic distances (100 to 200 nt), with a high level of sequence coverage in the IGS. In each case, we found that operon structures indicated by sequence data were confirmed by RT-PCR, thus validating the sequencing approach for operon identification.

One advantage of an unbiased approach is that we can directly test the current *B. anthracis* genome annotation by comparing it with global transcriptional data. First, we looked for transcription in regions that were originally annotated as genes but that had been subsequently removed from the genome. We found 36 loci that had been dropped from the genome but showed significant transcriptional activity and an additional 21 nonannotated regions with significant levels of transcriptional activity that were >250 nt from any known gene (Fig. 2B; see also the text in the supplemental material). Only two annotated genes revealed no transcriptional activity at all

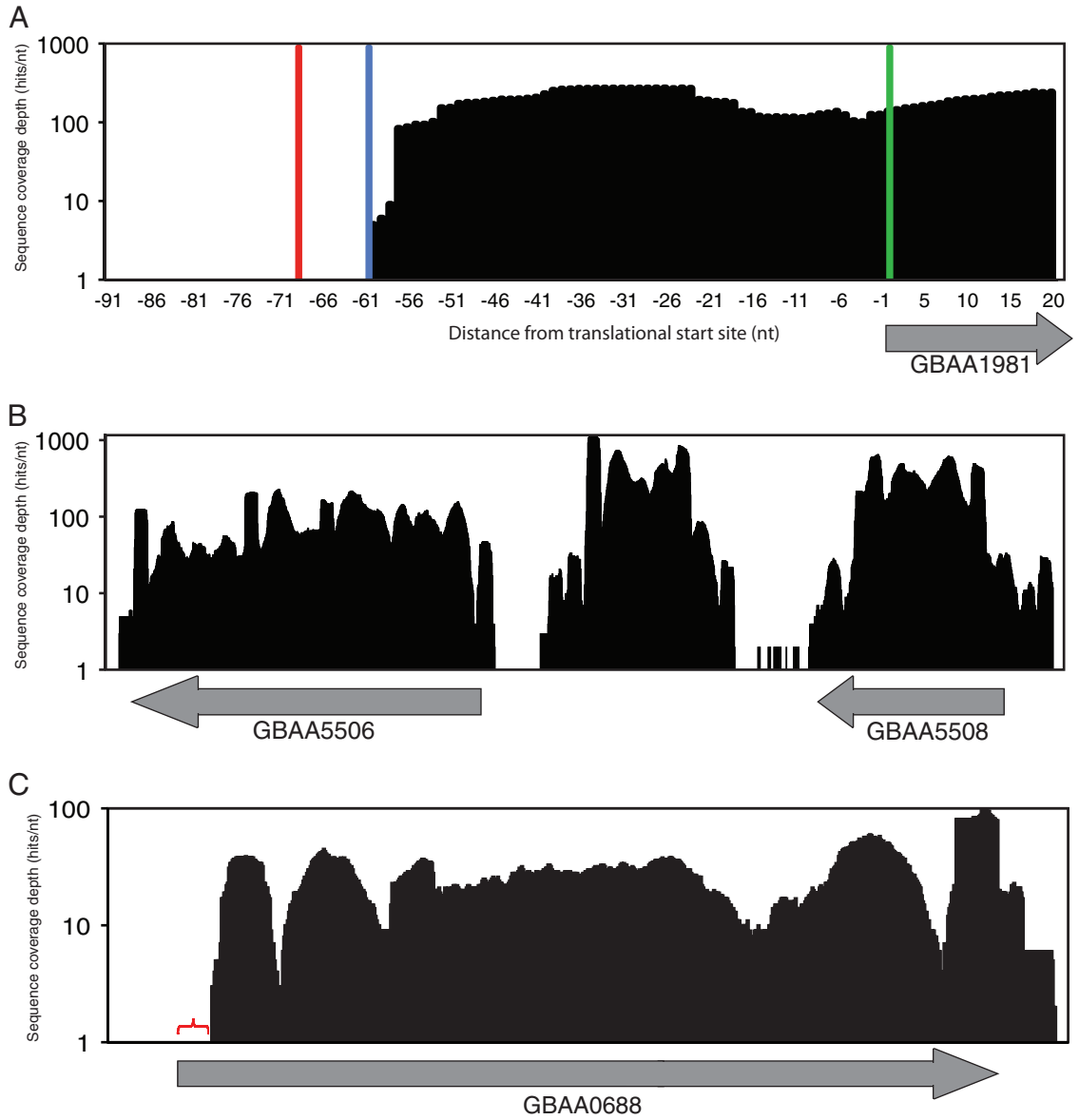

FIG. 2. Single-nucleotide resolution-sequencing data allow multiple mapping strategies to reveal global transcriptome composition. (A) Sequence coverage near the 5′ terminus of the *asbA* (GBAA1981) gene. The green line indicates the start codon, with the arrow beneath showing the direction of transcription. The blue and red lines indicate the transcript start site determined by sequence coverage and template-switching extension (TSx) data, respectively. (B) Sequence coverage in the GBAA5506-8 region of the *B. anthracis* chromosome, showing transcription across a region not included in the current genome annotation. Arrows beneath show the positions of the GBAA5506 and GBAA5508 loci. (C) Sequence coverage near the GBAA0688 locus, with a clear boundary inside the annotated gene (gap between annotated start codon and beginning of sequence coverage noted by red bracket).

(GBAA0736 and -3309); both are small (132 and 96 bp, respectively) and encode hypothetical proteins. Lastly, a persistent challenge in bacterial genome annotation is differentiating internal methionine codons from true translational start sites. We searched our coverage data for transcripts with 5′ ends starting downstream of the currently annotated translational start site and identified 11 genes whose start codons may have been incorrectly annotated (Fig. 2C; see also the text in the supplemental material).

We note that the RNA-Seq method has the potential to identify small regulatory RNAs as well, and we expect that our data may reflect the presence of these transcripts as well. However, a comprehensive sampling of small RNAs requires very different protocols for RNA isolation and library preparation, and we expect that some of the steps in the protocols used for this study, although essential for the experiments described here, may have biased our samples against small RNA sequences. We have therefore not included an analysis of these elements in this study, and a more complete analysis of noncoding RNAs in *B. anthracis* is currently under way in our laboratory.

Global transcript abundance and complexity in *B. anthracis*. In addition to providing a means of determining the physical

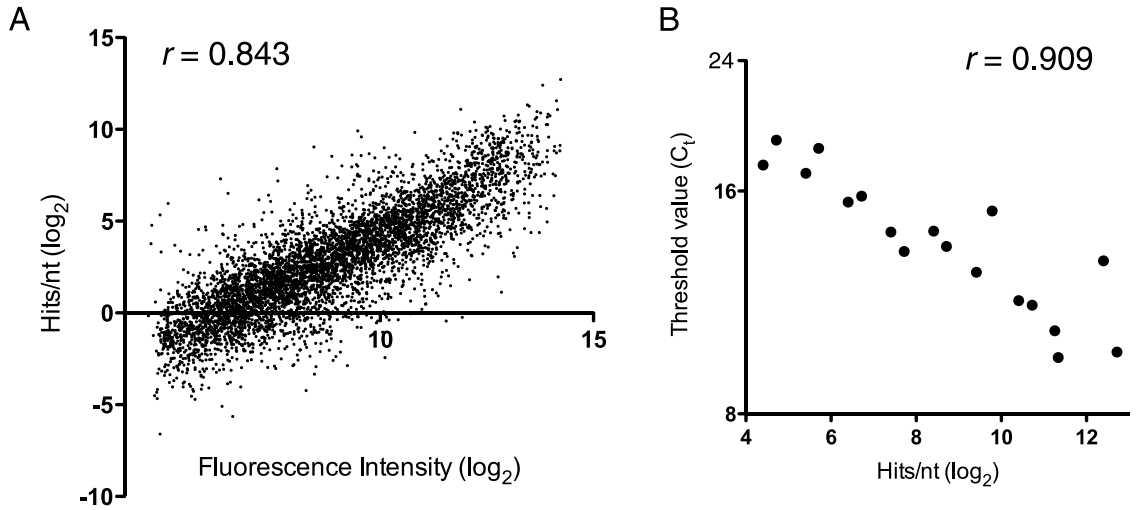

FIG. 3. Sequence coverage depth is quantitative and accurately reflects *B. anthracis* transcript abundance. (A) Comparison of sequence coverage depth and absolute Affymetrix GeneChip intensities. Shown is a plot of sequence coverage depth for sample 5 and raw microarray intensities for an equivalent sample (i.e., collected at the same point in the life cycle under the same growth conditions) analyzed previously and reported in reference 1. (B) Comparison of sequence coverage depth with SYBR green qRT-PCR data using RNA from sample 6.

structure of the *B. anthracis* transcriptome, the sequence data collected in our study present an opportunity to view gene expression in a way that is not biased by sequence-specific differences in hybridization efficiency. In order to minimize the effects of noise, we measured the expression level of each gene as an average coverage depth across the entire length of each gene, expressed in hits per nucleotide (a small number of genes [<100] that contained exact repeats and therefore did not allow for unambiguous mapping were not considered). As expected, these expression measurements showed a strong correlation with both absolute microarray intensities (1) and qPCR data (Fig. 3A and B), though we note that the dynamic range for the sequencing-based data was much greater ($\sim 2^{20}$ for sequence data versus $\sim 2^{10}$ for microarray data). This difference is consistent with the results of other RNA-Seq studies (14, 17, 23, 24) and helps explain why the correlation between array- and sequencing-based expression measurements is somewhat weaker for genes expressed at extremely high or low levels.

Conveniently, genome-wide studies of *B. anthracis* gene expression throughout its life cycle in vitro or during growth in various atmospheric conditions have been published recently (1, 2, 20), and we were able to use these data in confirming the validity of the approach used here. As seen in the representative genome segment shown in Fig. 4A, when we compared the sequencing-based expression measurements from samples 5 and 7 across the GBAA1979-88 region (after normalizing for total unambiguously mapped read count across the genome; see Materials and Methods), we observed that in late sporulation (sample 7), genes GBAA1979-80 appear to be transcribed at roughly the same level as in the log-phase sample (sample 5), while genes GBAA1981-6 and GBAA1987 are differentially expressed (roughly 10-fold down and 100-fold up during sporulation, respectively). When we compared these trends to the observations made using array data obtained from equivalent samples (i.e., from RNA isolated at the same point in the life cycle and under the same growth conditions), we found a pattern of differential expression that is essentially identical: genes GBAA1979-80 show an unchanged level of expression, expression of the GBAA1981-6 operon is 14-fold lower, and expression of the GBAA1987 locus is 140-fold higher during sporulation.

This level of agreement was typical across the genome; overall, the sequencing and array-based transcriptional profiles were highly consistent, and the global expression patterns that we observed in our sequencing data matched very closely with what has been reported by our group and others in previous microarray studies (1, 2, 20). This is consistent with several recent papers that have demonstrated that sequencing-based transcriptional profiling is at least as accurate and reproducible as current array-based methods (14, 23), and it seems clear that RNA-Seq as described here has the potential to be an extremely powerful tool for studying bacterial gene expression.

Apart from the advantages that we and others have reported—greater dynamic range, better accuracy and reproducibility, as well as greater flexibility (i.e., it requires no array design and construction and can be used for any microbe, even those for which no finished genome sequence exists)—it is also worth noting that the resolution provided by a sequencing-based approach allows us to directly visualize some of the more subtle elements of bacterial gene regulation. Figure 4B shows an example of this: here, in comparing coverage profiles of *B. anthracis* growing in O_2- and CO_2-rich environments, we observed a putative 8-gene operon in which the first gene (*ilvE-1*; GBAA1416) showed very little change in expression levels between the two growth conditions, while the downstream genes showed a significantly higher expression level in CO_2. This general trend was confirmed in array-based expression experiments (20) and seemed to suggest the presence of a structural element within the transcript (between GBAA1416 and GBAA1417) that plays a role in determining whether RNA polymerase is allowed to transcribe the remaining genes in the operon. Although very little experimental work has been done in characterizing such regulatory elements in *B. anthracis*,

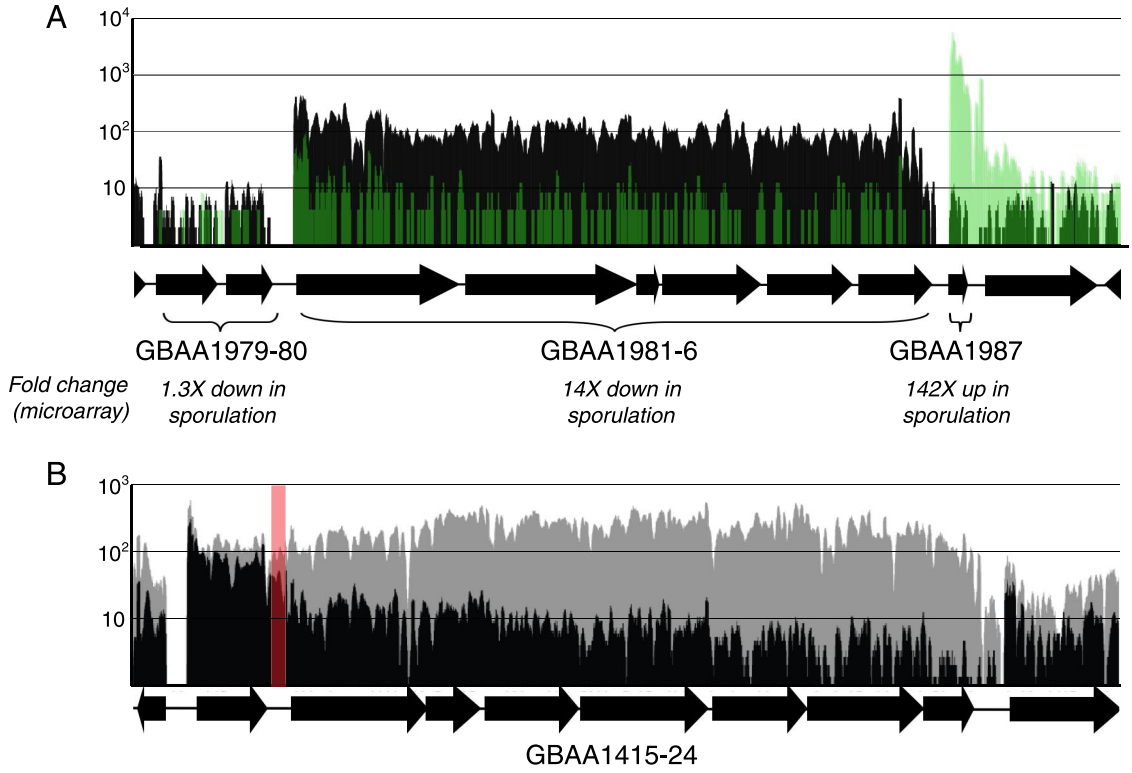

FIG. 4. Transcriptional profiling in *B. anthracis* using RNA-Seq. (A) Sequence coverage plots for samples 5 (mid-log phase; in black) and 7 (late sporulation; in green), normalized to the genome-wide total number of reads mapped unambiguously. The representative region shown is an ~11-kb genome segment surrounding the *asb* operon (GBAA1981-6), and specific genes are indicated. Below each gene or operon is the change in expression level that was measured by comparing microarray data from equivalent samples (previously described in reference 1) (ArrayExpress accession number E-MEXP-788). (B) Coverage plots for the ~12-kb region surrounding the *ilvE-1-leuD* operon (GBAA1416-23) from samples 5 (mid-log phase in air; in black) and 6 (late-log phase in CO_2; in gray). The T-box structural element predicted by Griffiths-Jones et al. (7) is indicated in red.

the creators of the Rfam database (7) used a bioinformatic approach to identify putative RNA structural elements in the *B. anthracis* genome, and their study predicted the presence of a T-box riboswitch directly between genes at GBAA1416 and GBAA1417. The putative T-box sequence is highly conserved among the *Bacilli* (8, 25), and it will be interesting to see if further experiments confirm the activity and function of this element. More broadly, it appears that sequencing-based transcriptional profiles provide an unusually informative view of RNA structural elements and their influence on transcription, and ongoing work in our lab is exploring this in more detail.

As noted above, one of the strengths of the RNA-Seq approach is that it is an inherently unbiased method, so unlike array-based methods, it allows for a rough assessment of each transcript's absolute abundance. With this in mind, we sought to measure the abundance of each transcript in the *B. anthracis* genome and compile a quantitative profile for the complete transcriptome under each growth condition sampled. As seen in Fig. 5, the absolute expression levels of genes in the *B. anthracis* genome follow a continuous distribution, with no obvious divisions into discrete classes expressed at high or low levels. Significantly, the overall shape and continuous nature of this distribution is essentially invariant, even when overall gene expression patterns differ greatly (see Fig. S5 in the supplemental material). This finding, combined with previous indica-

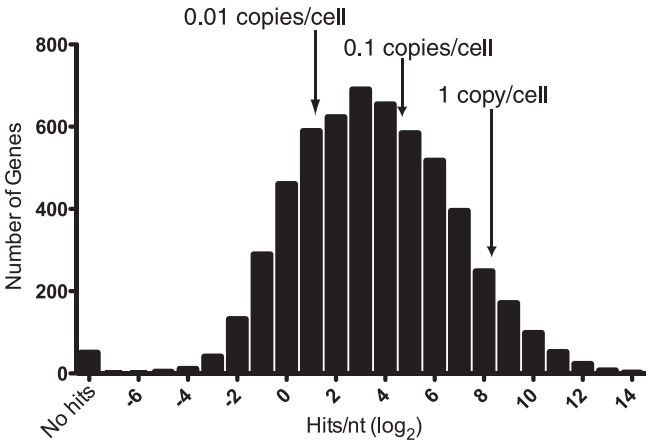

FIG. 5. The distribution of mRNA abundance in *B. anthracis*. The histogram shows the distribution of transcription levels (sequence coverage depth) for all genes in the *B. anthracis* genome in RNA sample 5. Arrows indicate the coverage depth expected (95% probability) for an mRNA molecule of average length present at 1 copy per cell ($2^{8.2–8.3}$), 0.1 copies per cell ($2^{4.7–5.0}$), and 0.01 copies per cell ($2^{1.3–1.9}$), based on the model described in the text in the supplemental material.

tions that *Escherichia coli* mRNA expression levels also follow a continuous distribution (9), implies that this feature of the transcriptome may be a general property of bacterial mRNA populations.

Interestingly, although the large virulence-associated plasmid pXO1, which is carried by the *B. anthracis* Sterne $34F_2$ strain used in this study, is presumed to be present at a higher copy number per cell than the chromosome itself, we observed that in every sample assayed the average expression level of genes located on pXO1 was lower than the corresponding average level of chromosomally carried genes (4). This difference was statistically significant in each sample (*P* values by Welch's *t* test ranged from 0.015 for sporulating cells in CO_2 to 4×10^{-32} for log-phase cells in air) and was most pronounced in samples collected during early or mid-log phase and in air rather than CO_2, which is consistent with previous studies showing that many genes on pXO1 are upregulated late in the life cycle and in the presence of CO_2 (2, 10, 12, 20). It is not yet clear what the implications of this trend may be for *B. anthracis* biology and pathogenesis or how common this pattern is in other bacteria, but we note that despite the differences in overall average expression levels, the transcripts derived from pXO1 exhibit an abundance distribution that is similar in shape and continuity to the chromosomal distribution shown in Fig. 5, which seems to support the idea that this distribution may be universal in bacteria.

The absence of discrete abundance classes has a number of implications for bacterial gene expression and regulation. First, it is clear that there is no obvious separation between genes that are expressed and those that are not. Rather, there are simply degrees of expression across a large range. Given this finding, we sought to give biological significance to these data by translating coverage depth into absolute mRNA abundance. Although precise conversion is impossible without internal standards, we can nevertheless make reasonable estimates based on a simple statistical model in which the sequence coverage for a typical transcript follows a Poisson distribution (see the text in the supplemental material for a detailed description). Assuming there are roughly 1,400 mRNA molecules per cell (11), our model predicts the approximate coverage depth that would be expected for mRNA molecules present at various levels of absolute abundance in a typical cell (Fig. 5). Although transcript abundance does not necessarily correspond to precise protein abundance, the predictions of the model make intuitive sense. For instance, we observed that in sample 5 (mid-log-phase growth), most of the sporulation-associated transcripts are present at a level of <0.01 copies per cell, consistent with observations by our group and others that sporulating cells are very rare but can be found in a log-phase culture (data not shown). Both our model and the mRNA abundance distribution we observed suggest that this sort of heterogeneity is more the rule than the exception and implies that many transcripts are present in only a fraction of the cells.

This is perhaps the most interesting implication of the transcript abundance distribution (Fig. 5; see also Fig. S5 in the supplemental material): that within a highly synchronized clonal culture (1), individual cells apparently show a great deal of diversity at the mRNA level. Along the same lines, several studies in recent years have shown that gene expression has an inherently stochastic component, which leads to transcriptional, translational, and subsequent phenotypic diversity among cells within a clonal population (5, 19, 21). Most of these studies have focused on the expression of a single gene, and our data provide a complementary view of this phenomenon by highlighting on a genome-wide scale the heterogeneity that is present in a seemingly homogeneous bacterial population.

In this study, we provided the first unbiased and comprehensive view of a bacterial transcriptome. We have defined in detail the RNA populations found in *B. anthracis* throughout its life cycle, and we have shown that our data can be used to map transcript boundaries and operon structure on a genome-wide scale and to identify previously unrecognized elements in the genome, thereby enhancing existing annotations. Further, the unbiased nature of our approach allowed us to view the transcriptome from an entirely new perspective and observe global trends in transcription that provide insights into stochasticity and diversity within bacterial populations.

ACKNOWLEDGMENTS

We thank Charles Cochran of Applied Biosystems for advice and assistance in collecting SOLiD sequence data, Scott Kuersten for helpful advice regarding sample preparation, and members of the Bergman lab for useful discussions.

This work was supported by DHHS contract N266200400059C/N01-AI-40059 (N.H.B.) and by a New Opportunities award from the Southeast RCE for Biodefense and Emerging Infectious Diseases.

REFERENCES

1. **Bergman, N. H., E. C. Anderson, E. E. Swenson, M. M. Niemeyer, A. D. Miyoshi, and P. C. Hanna.** 2006. Transcriptional profiling of the *Bacillus anthracis* life cycle in vitro and an implied model for regulation of spore formation. J. Bacteriol. **188:**6092–6100.
2. **Bourgogne, A., M. Drysdale, S. G. Hilsenbeck, S. N. Peterson, and T. M. Koehler.** 2003. Global effects of virulence gene regulators in a *Bacillus anthracis* strain with both virulence plasmids. Infect. Immun. **71:**2736–2743.
3. **Cendrowski, S. R.** 2004. Role of the asb operon in Bacillus anthracis pathogenesis. Ph.D. dissertation. University of Michigan, Ann Arbor.
4. **Coker, P. R., K. L. Smith, P. F. Fellows, G. Rybachuck, K. G. Kousoulas, and M. E. Hugh-Jones.** 2003. *Bacillus anthracis* virulence in guinea pigs vaccinated with Anthrax Vaccine Adsorbed is linked to plasmid quantities and clonality. J. Clin. Microbiol. **41:**1212–1218.
5. **Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain.** 2002. Stochastic gene expression in a single cell. Science **297:**1183–1186.
6. **Gama-Castro, S., V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A. M. Huerta, L. Trevino-Quintanilla, and J. Collado-Vides.** 2008. RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res. **36:**D120–D124.
7. **Griffiths-Jones, S., S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman.** 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. **33:**D121–D124.
8. **Gutiérrez-Preciado, A., T. M. Henkin, F. J. Grundy, C. Yanofsky, and E. Merino.** 2009. Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. Microbiol. Mol. Biol. Rev. **73:**36–61.
9. **Hereford, L. M., and M. Rosbash.** 1977. Number and distribution of polyadenylated RNA sequences in yeast. Cell **10:**453–462.
10. **Hoffmaster, A. R., and T. M. Koehler.** 1997. The anthrax toxin activator gene *atxA* is associated with CO_2-enhanced non-toxin gene expression in *Bacillus anthracis*. Infect. Immun. **65:**3091–3099.
11. **Ingraham, J. L., O. Maaløe, and F. C. Neidhardt.** 1983. Growth of the bacterial cell. Sinauer Associates, Sunderland, MA.
12. **Koehler, T. M., Z. Dai, and M. Kaufman-Yarbray.** 1994. Regulation of the *Bacillus anthracis* protective antigen gene: CO_2 and a *trans*-acting element activate transcription from one of two promoters. J. Bacteriol. **176:**586–595.
13. **Li, R., Y. Li, K. Kristiansen, and J. Wang.** 2008. SOAP: short oligonucleotide alignment program. Bioinformatics **24:**713–714.
14. **Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad.** 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. **18:**1509–1517.

15. **Matz, M. V., N. O. Alieva, A. Chenchik, and S. Lukyanov.** 2003. Amplification of cDNA ends using PCR suppression effect and step-out PCR. Methods Mol. Biol. **221:**41–49.
16. **McGrath, P. T., H. Lee, L. Zhang, A. A. Iniesta, A. K. Hottes, M. H. Tan, N. J. Hillson, P. Hu, L. Shapiro, and H. H. McAdams.** 2007. High-throughput identification of transcription start sites, conserved promoter motifs and predicted regulons. Nat. Biotechnol. **25:**584–592.
17. **Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder.** 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320:**1344–1349.
18. **Ondov, B., A. Varadarajan, K. D. Passalacqua, and N. H. Bergman.** 2008. Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. Bioinformatics **24:**2776–2777.
19. **Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden.** 2002. Regulation of noise in the expression of a single gene. Nat. Genet. **31:**69–73.
20. **Passalacqua, K. D., A. Varadarajan, B. Byrd, and N. H. Bergman.** 19 March 2009. Comparative transcriptional profiling of *Bacillus cereus* sensu lato strains during growth in CO_2 bicarbonate and aerobic atmospheres. PLoS One **4:**e4904. [Epub ahead of print.]
21. **Raj, A., and A. van Oudenaarden.** 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell **135:**216–226.
22. **Rocha, E. P.** 2004. Order and disorder in bacterial genomes. Curr. Opin. Microbiol. **7:**519–527.
23. **'t Hoen, P. A., Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. Vossen, R. X. de Menezes, J. M. Boer, G. J. van Ommen, and J. T. den Dunnen.** 15 October 2008. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. **36:**e141. doi:10.1093/nar/gkn705.
24. **Wilhelm, B. T., S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, and J. Bahler.** 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature **453:**1239–1243.
25. **Winkler, W. C., F. J. Grundy, B. A. Murphy, and T. M. Henkin.** 2001. The GA motif: an RNA element common to bacterial antitermination systems, rRNA, and eukaryotic RNAs. RNA **7:**1165–1172.