The Regulatory RNAs of Bacillus subtilis
Mars, Ruben
Chapter 2

Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis


*These authors contributed equally to this work.

Published in Science. 2012, 335, 1103-1106
Abstract

Bacteria adapt to environmental stimuli by adjusting their transcriptomes in a complex manner, the full potential of which has yet to be established for any individual bacterial species. Here, we report the transcriptomes of *Bacillus subtilis* exposed to a wide range of environmental and nutritional conditions that the organism might encounter in nature. We comprehensively mapped transcription units (TUs) and grouped 2935 promoters into regulons controlled by various RNA polymerase sigma factors, accounting for ~66% of the observed variance in transcriptional activity. This global classification of promoters and detailed description of TUs revealed that a large proportion of the detected antisense RNAs arose from potentially spurious transcription initiation by alternative sigma factors and from imperfect control of transcription termination.
Introduction

Bacterial transcriptomes are surprisingly complex (1, 2, 3, 4, 5) and include diverse and abundant small RNAs and antisense RNAs (asRNAs). Because only a small number of environmental conditions have been investigated, the full extent of transcriptome complexity remains to be established for a single bacterial species. This prompted us to undertake a systematic and quantitative exploration of transcriptome changes in \textit{Bacillus subtilis}, whose natural habitat, the soil, is subject to severe environmental fluctuations (6). \textit{B. subtilis} is also a laboratory model for Gram-positive bacteria and is grown in industrial scale fermentors for the production of enzymes and vitamins. We selected 104 conditions covering the use of various nutrients, aerobic and anaerobic growth, the development of motility, biofilm formation, adaptation to diverse stresses, high–cell density fermentation, and hallmark adaptations of \textit{B. subtilis}: the development of competence for genetic transformation, cell differentiation to form resistant spores, and germination from such spores.

In our labs, we grew a prototrophic strain under the selected conditions and hybridized 269 RNA samples to tiled microarrays with a resolution of 22 bases (supporting online material (SOM) 1 and Table S1) (7). We then estimated strand-specific RNA signals from the raw data (Figure 1A) (8) and computed an aggregated expression index for all transcribed regions (SOM 2 and Table S2). The expression threshold was stringently set to reduce the number of potentially false short RNA features detected. Of the previously annotated coding sequences (CDSs), only 186 (4.4%) were not expressed under any condition. Most of these CDSs were of unknown function and predicted to originate from horizontal transfer (SOM 3 and Table S3). The 30% of the CDSs most highly expressed under each condition were defined as “highly expressed” (SOM 3). Eighty-five percent of all CDSs were highly expressed in one or more conditions (Figure S3A), but only ~3% (144) of all CDSs were highly expressed under all conditions, indicating that most \textit{B. subtilis} genes are differentially expressed. Genes in the latter group encode proteins with essential functions (9) and enzymes involved in glycolysis, iron sulfur metabolism, and detoxification pathways (Table S3).

For each of the 269 profiles, we used the repertoire of high-confidence 5’ and 3’ mRNA ends to delineate the transcribed regions, which we derived directly from the positions of abrupt increases and decreases in RNA abundance, hereafter called up- and downshifts, respectively (Figure 1A and SOM 2). More than 85% of previously known transcriptional start sites (10, 11) mapped within ~12 base pairs (bp) of an upshift (SOM 4), indicating that most mRNAs have unprocessed 5’ ends. Therefore, most of the 3242 detected upshifts appear to correspond to genuine promoters, increasing ~threefold the number of promoters mapped in \textit{B. subtilis}. Downstream of these promoters, 3000 transcription units (TUs) were defined and then decomposed into previously annotated genes and newly identified RNAs (SOM 2 and Tables S4 and S5). This analysis revealed that ~46% of all annotated CDSs can be transcribed from more than one promoter. After manual validation, we classified 1583 previously unannotated RNAs (S1 to S1583) according to their structural relationships with neighboring annotated CDSs (Table 1). These RNAs considerably expand the registry of potential regulatory RNAs (SOM 5 and table S6) and increase the total number of potential genes in \textit{B. subtilis} by 11% (Table 1). The condition dependent experimental annotation can be queried at www.basysbio.eu/bsubtranscriptome/seb.

The 3’ ends of most mRNAs were associated with predicted intrinsic terminators (that is, an RNA hairpin followed by a U-track) at which RNA polymerase (RNAP) dissociates without the need for auxiliary proteins (SOM 6 and Table S7). A subgroup of 174 TUs generated asRNAs due to the lack of a termination site and readthrough transcription (categories Indep-NT, 3’NT, and 3’PT in Table 1). These mRNAs could extend up to 3.4 kb beyond the 3’ boundaries of their cognate CDS with a gradual decrease in signal intensity (Figure 1B). Overall, 13% of \textit{B. subtilis}
CDSs are overlapped by asRNAs (Table S11) that can potentially act to regulate their cognate sense mRNAs at the transcriptional, RNA stability, or translational levels (12). In a mutant strain lacking the termination factor Rho, which is not required for growth of *B. subtilis* in a rich medium (13), the mRNA extensions reached up to 12 kb (average ~2.8 kb) (Figure 1C and Table S9). Without Rho, additional asRNAs were formed by extension of a subset of TUs, many of which have only partially efficient intrinsic terminators (Figure 1C and SOM 7), indicating that Rho is a general inhibitor of antisense transcription.

The relations among the expression profiles revealed highly correlated changes in expression (Figure 1D and SOM 8) and provided functional insights (SOM 9). To elucidate the global transcriptional regulatory network, we used the mRNA signal downstream of each of the 3242 detected upshifts to estimate pairwise correlations between promoter activities (SOM 10). The results are summarized in the form of a tree created by hierarchical clustering (Figure 2A). In this tree built independently of DNA sequence information, the DNA binding sites for different RNAP sigma subunits (11) are clustered (Figure 2B), highlighting the prominent role of sigma factor–mediated promoter recognition in differential gene expression.

For *de novo* genome-wide identification of sigma factor regulons, we developed an unsupervised algorithm combining information from the DNA sequences near the upshifts and
Table 1. Summary of mRNA features.
The 1583 previously unidentified RNA features were categorized according to size (columns) and their structural relationships with annotated genes (SOM 2): 5' and 3' regions; Indep, RNA independent of previously annotated features; Inter, mRNA segment lying between two genes with distinct promoters; Intra, portion of a polycistronic mRNA from a single promoter. The 3' RNAs comprised three subgroups: 3' untranslated regions (3'UTR) and 3' extended transcripts with either no termination (3'NT) or partial termination (3'PT). Similarly, the Indep RNAs displayed either termination at a site (Indep) or no termination (Indep-NT) (Figure 1B). L, length; nt, nucleotides; n, number of features. Pred.; predicted.

<table>
<thead>
<tr>
<th>Features</th>
<th>50 ≤ L ≤ 150 nt</th>
<th>L ≥ 150 nt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>asRNA*</td>
</tr>
<tr>
<td>5'</td>
<td>462</td>
<td>5</td>
</tr>
<tr>
<td>3'UTR</td>
<td>64</td>
<td>5</td>
</tr>
<tr>
<td>3'NT</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3'PT</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Indep</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Indep-NT</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Inter</td>
<td>182</td>
<td>3</td>
</tr>
<tr>
<td>Intra</td>
<td>132</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>865</td>
<td>17</td>
</tr>
</tbody>
</table>

*asRNAs are segments overlapping ≥100 bp (or 50% for short segments) with the sense RNA. †Putative unannotated CDS (≥10 amino acids) predicted by the software SHOW with a confidence score above 90% (25). ‡Previously unidentified genes include all Indep, Indep-NT, asRNAs, and predicted CDSs. Four genes encoded both asRNAs and potential CDSs. The RNA features are listed in Table S5.

Figure 2. Classification of promoters.
A) Hierarchical clustering tree summarizing the correlations (x axis) between promoter pairs arranged on the y axis. Detailed information on each promoter and its neighbors can be found in Table S4. B) Each promoter is shown as a horizontal bar that is colored according to the distinct clusters revealed by the unsupervised identification of bipartite motifs. Color coding of clusters is as displayed in Figure 2C, and promoters with low-confidence cluster assignment are shown in gray. For each cluster, the defined consensus motif is labeled with the sigma factor name (letter) and motif number (as in Table S4). C) Estimated activity of clusters (y axis, log2 scale) across the conditions that are arranged on the x axis according to the “shortest tour” (SOM 8 and Figure S4). D) Decomposition of the total variance of promoter activity (SOM 10 and Table S12). Unexplained variance is shown in gray.
Condition-dependent transcriptome from the clustering of promoters to model the bipartite degenerate motifs recognized by the sigma-containing RNAP (SOM 10). This approach is an alternative to the targeted search for motifs in prespecified clusters (14). A predicted sigma factor binding site could be assigned to the majority (2935/3242) of promoters identified (Figure 2B and Table S12). The promoters expected to bind SigA formed six clusters reflecting substantial motif diversity (15), and SigA bound in vivo to all six motifs (SOM 11). The promoters recognized by the other sigma factors formed either distinct clusters, each with a distinctive DNA binding motif (SigB, SigD, SigH, SigI, SigK, and SigL), or clusters that grouped several sigma factors that displayed similar DNA binding motifs (Figure 2B and SOM 12). This de novo classification is consistent with the cross-recognition observed between the sporulation-specific SigE, SigF, and SigG factors (16, 17) and the extracellular function sigma-factors SigM, SigW, SigX, and SigY (18, 19).

The comprehensive definition of regulons allows quantification of the contribution of each sigma factor to transcriptome plasticity in B. subtilis. The global activity of a single motif cluster was estimated as the average transcription signal for all the promoters of the cluster (Figure 2C). Assuming a simple linear relation between individual promoters and cluster activities, we computed the fraction of the transcriptome variance explained by each cluster motif (SOM 10). Overall, ~66% of the observed variance was attributed to variations in the activity of sigma factors mainly associated with sporulation and responses to stress, processes that are essential for adaptation and survival in nature (Figure 2D and Table S12). The proportion of variation attributed to sigma factor activities varied considerably. The highest values (>0.75) were obtained for SigB, SigD, and the sporulation sigma factors. In contrast, the lowest value (0.24) was obtained for SigA, suggesting that condition-specific regulation of SigA-dependent promoters relies mostly on other transcription factors (TFs). Our classification captured biologically meaningful regulatory information on TF function, despite the multilevel organization of the TF regulatory network (SOM 13) (20). It also provides leads for further functional analyses and biotechnological applications (SOM 14).

AsRNAs exhibited diverse condition-dependent expression profiles (Figure 3). Indeed, AsRNAs are more often initiated from non-SigA promoters (48%) than protein-coding RNAs (26%) (SOM 15 and Table S14). AsRNAs also tend to be expressed in lower amounts than protein-coding RNAs (Figure S19), and expression of sense and asRNA pairs displays mutual exclusion more often than coexpression (47% negative correlation versus 30% positive correlation). The predominance of negative correlation results primarily from the differential expression of sense and antisense transcripts from SigA and non-SigA promoters, respectively (SOM 15 and Figure S20). Transcription extending beyond the boundaries of CDSs often generates asRNAs

![Figure 3. Patterns of natural antisense transcription across conditions.](image_url)

A) Variation of the number of highly expressed (upper 30%) asRNAs across the conditions (x axis). B) Heat map representation: asRNAs are ordered as in Table S11 and are colored according to their expression levels.
(Figure 1B). Importantly, inefficient termination of a subset of TUs is globally counteracted by the action of Rho, consistent with its role in matching transcription to translational needs (21, 22, 23). Altogether, ~80% of asRNAs result from transcription initiated by alternative sigma factors or from imperfect transcription termination (Table S14). We propose that many asRNAs arise from spurious transcriptional events, which are more prevalent when driven by alternative sigma factors. Supporting this hypothesis, the sigma factor binding sites appear to be less evolutionarily conserved in promoters driving potentially spurious asRNAs than in those driving protein-coding RNAs (SOM 15 and Figure S23). Nevertheless, some asRNAs might have important biological functions (12). We detected asRNAs with established or potential functions (Table S6), the latter including the asRNAs expressed in amounts sufficient to regulate their target genes (SOM 15 and Figure S19). In conclusion, our quantitative analysis of the condition dependent transcriptome defined the contribution of the sigma factor complement to B. subtilis transcriptome plasticity. It revealed that asRNAs generated by inefficient control of transcriptional events might be a drawback of this plasticity, though they might contribute to the creation of previously unknown regulatory functions.

Acknowledgments

The data are deposited in the Gene Expression Omnibus (GEO) database (accession numbers GSE27219, GSE27303, GSE27419, and GSE27652), and are tabulated in the SOM and at www.baysbio.eu/bsubtranscriptome. This Web page provides links to the data archived in the GEO and Openbis databases. This study was supported by the European Commission–funded BaSysBio project (LSHG-CT-2006-037469), the SysMO network funded by the German Federal Ministry of Education and Research (0313978B), Department of Education, Science and Training (CG110055), and the National Health and Medical Research Council (455646). We are grateful to N. Pelé, B. Giovani, C. Sautot, and B. Diep for management support. The authors declare no competing interests.

Supplementary Material

www.sciencemag.org/cgi/content/full/335/6072/1103/DC1
SOM Text (SOM 1 to 15)
Figs. S1 to S23
Tables S1 to S14
References (26–102)
Database S1
Condition-dependent transcriptome

Treatment of raw data

For each probe, the probe affinity covariate was computed from 4 genomic DNA hybridizations (8). We also computed SeqS, a single value reflecting the amount of sequence similarity susceptible of causing cross-hybridization as described in (26). To reconstruct the underlying transcription signal from the raw data, each hybridization was processed with our software “hmmtiling” (8) using options “-nS 100 -itermax 100 -nexpl 10 -epsexplo 1 -eta flat -randompar T”. With these settings, the data for probes with SeqS>1.9 were discarded (SeqS≥2 for probes with multiple exact matches). From the “hmmtiling” software output, the following parameters were used to reconstruct the mRNA abundance signal and to map the up- and down-shifts: ES[i] the expected value of the underlying signal at probe i, CIminS[i] the lower end of the 95% credibility interval for the underlying signal, VitM[i] indicating signal shifts and drifts (VitM ≠ 0) in the most likely signal trajectory and PM1up[i] and PM1down[i] the probabilities of upward and downward signal shift.

The microarray design (7) was based on the GenBank record AL009126.2, prior to the publication of AL009126.3. We used the Mauve global genome aligner (27) to map the probe positions from the v2 sequence 15 onto the v3 sequence.

Generating transcription parts lists

Transcript ends (high-confidence lists of signal up-shifts and down-shifts)

For each array, the positions of the up-shifts were identified in the most likely trajectory (VitM) and a confidence score for the up-shift at probe i (signal increase between i-1 and i) was computed as the expected number of up-shifts in a window of two adjacent probes (the maximum of PM1up[i-1]+PM1up[i] and PM1up[i]+PM1up[i+1]). Up-shifts with confidence above 0.99 served to establish a catalog of the high-confidence up-shifts (transcript 5'-ends) detected in the 269 hybridizations. Up-shift positions were defined as the coordinates of the 5’-end of the probe i. When up-shifts at adjacent probes were detected in different hybridizations they were considered as the same transcript end, and the consensus position for the up-shift was computed as the average position. The min position and the max position were also recorded. Exactly the same procedure was applied to establish the catalog of high confidence down-shifts (transcript 3’-ends).

Transcribed regions

To detect transcribed regions (TRs) outside of the annotated CDSs and RNA genes, the locally reconstructed signal was compared to the median for the whole chromosome (median ES). The threshold values to define the TRs were determined empirically as described in SOM3. We established a list of chromosome segments where the lower boundary of the 95% credibility interval (CIminS) for the inferred transcription level exceeds 10x the chromosome median in at least one hybridization (90.8% of the annotated CDSs were at least partly overlapping these regions). In parallel, we established a repertoire of probable 5’ and 3’ transcript ends corresponding to the position of high-confidence shifts across the 269 profiles (see above). TRs were then decomposed into segments according to the annotation and to the repertoires of transcript ends. To avoid undesirable effects of sharp boundary truncation at a determined cut-off, which would result in artifactual splitting of TRs, regions were extended on both sides according to a relaxed but still stringent 5x cut-off (see SOM3, 95.6% of the annotated CDSs had an expression index 5-fold higher than the background in at least one hybridization) and connected regions were merged. A final list of TRs containing the annotated CDSs and RNA
genes and the novel TRs (size > 50 nt) was obtained after automatic and manual curation (see below, and Table S2 and S5). For each TR, the expression level is summarized by a single value computed as the median ES for probes within the TR having a perfect unique match on AL009126.3. The whole data set was deposited in GEO (GSE27219).

**Transcription units**

A transcription unit (TU) contains one or more structural genes which are transcribed from a single promoter into a mono- or poly-cistronic mRNA. Thus, in a given condition, the expression levels of genes in a TU remain largely constant without being interrupted by a “low signal” gap, assuming no differential mRNA degradation (28). However, across the many diverse conditions used in this study, the criterion of a constant expression level is not sufficient to define TUs. Thus, we used a more operational definition for the TUs. All the individual transcription profiles were analyzed to identify the extent of the transcribed region associated with each signal up-shift (5’-end). After examining all the hybridizations, a TU was defined as the maximal extent of the region expressed (CIminS > 5x chromosome median) after a signal increase (VitM ≠ 0) at a position of a high-confidence up-shift (transcript 5’-end). According to this definition, the expression levels of genes within a TU are not necessarily constant. The level of mRNA can decrease if the TU contains down-shifts, (probable 3’-ends) provided that the signal remains above 5x chromosome median in at least one hybridization. The observed levels of mRNA can also result from the activities of several promoters generating overlapping transcripts, and two or more successive up-shifts can therefore be found without being interrupted by a “low signal” gap. Detecting the activity of the downstream promoter necessitates however that its activity is high enough to add to the upstream signal. Downstream of a second promoter associated with a signal increase (VitM ≠ 0) there is no direct information on the amount of signal that should be attributed to each upstream promoters, and the extension of the upstream TU is considered potential. When two promoters generated overlapping mRNAs, the upstream promoter was associated with a potential TU extended by the size of the overlapping transcribed region. Thus, two sets of predicted TUs were built, providing an upper and a lower boundary on the extent of each actual TU. For the upper boundary (TUlong), we considered the transcripts associated to successive upshifts as overlapping, whereas, for the lower boundary (TUshort), only to the last up-shift of the considered TU is taken into account. Complete lists of up-shifts and of predicted TUs are available in Table S4. Importantly, the total number 3242 up-shifts detected is probably underestimated since downstream promoters are detected only if their activity is sufficiently high to increase the mRNA level. Importantly, due to the stringent criteria used to define the transcribed regions, 242 of the 3242 high confidence up-shifts were not associated with a TU. For example, a high confidence promoter is identified upstream of the glnT and glnA genes (U175, Table S4) but the level of expression never reached the cut-off (above 5x chromosome median in at least one hybridization) under the conditions tested (Table S3, see also http://www.basysbio.eu/subtranscriptome/seb).

**Curation and classification of RNA features**

The procedure described for establishing the TR parts list led to the identification of 1836 new regions associated with expression level measures. After the automatic and manual curation, 1577 TRs were kept, 244 discarded and 15 merged into 6 new regions resulting in a final list of 1583 TRs (1836-244-15+6) numbered S1-S1583 according to their chromosomal position (Table S5).

The curation procedure started by the automatic generation of a list of 268 TRs proposed to be discarded based on expression measures across hybridizations not exhibiting the
Condition-dependent transcriptome: typical biological condition effect observed for annotated features (Figure S1). For each region, this effect was assessed by fitting a linear model on quantile-normalized data and examining the p-value for rejection of the null hypothesis of an expression level independent of biological conditions (regions with high variation between replicates compared to variation between biological conditions received higher p-value). While a high signal can cause saturation and thus can decrease the power to reject the null hypothesis it would generally be considered to increase the confidence in the TRs. Thus, the TRs having a median difference between the expression measure and background across hybridizations above 10x were not included in the list of 268 detected TRs to be discarded. After systematic visual examination of all TRs along the chromosome by at least 3 expert biologists, 42 TRs were rescued from the list of 268 TRs to be discarded and 18 not initially in this list were discarded (244=268+18-42).

The classification of the new TRs into the transcriptional context categories: 5', 3'UTR, 3'PT, 3'NT, Indep, Indep-NT, Inter and Intra (Table 1) was carried out automatically and then all new TRs were checked manually. Depending on the presence of an element of the down-shifts parts list (transcript 3'-ends), we further distinguished 3'NT from 3'UTR and Indep-NT from Indep. The categories were defined as described below, and the color coding used in Figure 1 and in the B. subtilis Expression data browser (http://www.basysbio.eu/bsubtranscriptome/seb) is indicated in parentheses:

- [5'] - 5' segment (green)
- [Intra] - transcribed segment lying between two genes under the control of the same promoter (dark blue)
- [Inter] - a transcribed segment between two genes with distinct promoters (light blue)
- [3'UTR] - 3' untranslated segment ending with a signal down-shift (red)
- [3'NT] - 3' transcribed segment lacking a site for transcription termination (no abrupt signal down-shift) and exhibiting a 3' extended mRNA with a slow decrease in the signal intensity (orange)
- [3'PT] - 3’ transcribed segment resulting from a partial (incomplete) termination of transcription (old yellow)
- [Indep] - Independent segments transcribed from their own promoter and ending with a down-shift (black)
- [Indep-NT] - Independent segments transcribed from their own promoter that do not end with a down-shift but exhibit a 3’ extended mRNA with a slow decrease in the signal intensity (black).

The new mRNA features overlapping GenBank features on the opposite strand were categorized antisense (AS) when the overlapping region was either ≥100 bp or, for regions shorter than 200 bp, when it accounted for ≥50% of the length of the new mRNA feature or the GenBank feature. GenBank features fulfilling these conditions were called antisense targets. Note that with this definition, all asRNAs have at least one antisense target, and reciprocally, all antisense targets have at least one asRNA.

(SOM 5) New RNA features expand the registry of potential regulatory RNAs

The comprehensive (though not exhaustive) condition-dependent experimental annotation of the TUs in the B. subtilis genome comprises 4 major tables providing the expression levels for all features in the 269 mRNA profiling experiments (Table S2), the list of promoter up-shifts with cluster information and TU definition (Table S4), a summary of information for each RNA feature (Table S5), and a list of high confidence downshifts and their classification (Table S7). The annotation can also be directly queried at http://www.basysbio.eu/bsubtranscriptome/seb. In this section, we show that our study delivered a compendium of valuable functional
information about the condition-dependent expression of genes, the promoters associated to
them, and about the 5’ and 3’ mRNA ends. The comparative analyses with previous studies are
summarized in Table S6.

New genes from expressed regions

The features of the most recent genome annotation (version 3) cover CDSs (4256), tRNAs (86),
rRNAs (30), cis-acting 5’ UTRs (57) involved in transcriptional and translational regulations,
and some RNA genes (5) universally conserved among bacteria (29). Here, we report 512
newly discovered genes (Table 1) corresponding to: i) all previously un-annotated RNA features
expressed from their own promoters (Indep, Indep-NT in Table 1, SOM 2), ii) the antisense
RNAs, and iii) the predicted CDS. As there is some overlap between these categories, the feature
type was assigned with the following priority: CDS > RNA gene > antisense RNA (Table S6.1).
The length distribution for each category is plotted Figure S6. For the 44 new CDSs detected
(Table 1), chromosome coordinates and CDS size are indicated (Table S6.1). Interestingly,
332/512 of the new genes encode antisense RNAs (Table S6.1, SOM 10).

Comparison of new RNA features from three high resolution approaches

We compared the 1583 RNA features (Table S5) with the results from recent studies by
Rasmussen et al. (7), also based on DNA tiling microarrays, and by Irnov et al. (10), based on a
deep sequencing approach of the primary (nascent and unprocessed) transcripts. Importantly,
the transcriptome was analyzed under 104 conditions in our study, under two conditions
(exponential phase in M9 minimal medium and in rich LB medium) in (7), and in one condition
(stationary phase in minimal medium with glucose) in (10).

Overall, we detected 79% and 50% of the new RNA features and asRNAs, respectively,
identified in (7). The corresponding ratios were 93% and 69% for the new RNA features and
asRNAs identified in (10). The new RNA features and asRNAs detected in at least two studies are
described in Table S6.2 and Table S6.3, respectively.

The study by Rasmussen et al. (2009) and ours used different stringency thresholds to
distinguish expressed from not expressed regions. Indeed, our threshold corresponds to a signaltimes above the background, therefore excluding a higher proportion of the background
noise inherent to the tiling array technology. This has an impact especially for the detection of
antisense signals. From a visual inspection of our transcriptome profiles, only four sense RNA
signals identified by Rasmussen and that we rejected had a clear expression profile that did not
reach our threshold (ncr9, 35, 43, 70), whereas the other RNA features identified by Rasmussen
were not distinguishable from the background in our data. Similarly, only five antisense signals
from (7) that we rejected displayed a clear below-threshold expression (shd12, 52, 61, 88, 117),
the remaining "shd" features being indistinguishable from the background. A much larger
fraction of the RNA features identified by (10) were detected in our study, suggesting a lower
level of noise in RNA-seq approaches. Nevertheless, our analysis of the transcriptome across
diverse conditions revealed a wealth of novel RNA features not detected in these two recent
studies.

Experimental evidence for RNA genes in B. subtilis

Five RNA housekeeping genes, present in the most recent version of the B. subtilis genome
annotation (29), are universally represented in the bacterial kingdom: small cytoplasmic RNA
(Scr), 6S RNAs (BsrA, BsrB), RNase P ribozyme (RnpB), and finally, tmRNA (SsrA). All of them
were identified in Rasmussen, Irnov and our study (Table S6.4).

By analyzing the literature, we found experimental evidence (5’ end mapping and
Condition-dependent transcriptome Northern blot analysis) for 15 additional RNA genes in *B. subtilis*. Among these, BsrC (30) and SurC (31) were not observed in our study, as well as those of Rasmussen and Irnov (Table S6.4). Two other RNA genes, SR1 (32) and BsrI (30) were assigned to CDSs when the genome annotation was upgraded from version 2 to version 3. Yet, the SR1 RNA acts directly to regulate the translation of the *ahrC* transcriptional activator (33), while its encoded peptide acts on the expression of the *gapA* operon (34). Of note, RatA (S976) is a small untranslated asRNA which overlaps 75 nucleotides with the gene encoding the toxic peptide TxpA (formerly YqdB), and controls *txpA* mRNA expression (35). Although S976 was not classified as asRNA because the antisense overlap length was <100 nt, it was detected and annotated in Table S5 (overlap length of 58 nt with our stringent settings). In addition, we predicted 2 new CDS in BsrG (27 amino acids) and BsrH (38 amino acids) with a confidence prediction of 94% and 99% respectively. Overall, 18 of the 20 known small RNAs were detected in our study.

Validation of long 5’ leader regions

Some 676 5’ leader regions ranging from 50 to 4226 bases in size were annotated, including regions with multiple promoters. They provide an unbiased list of condition-dependent 5’ leader regions that can be mined to identify new cis-acting regulatory RNAs (Table S5). Irnov et al (10) provided a selected list of 40 candidates for long 5’ leader regions matching two criteria: the RNA-seq reads start within an intergenic region and overlap with the downstream CDS.

Figure S6: Length distributions of the different types of RNA features

The length distribution of the 8 types of RNA features describing different transcriptional context are represented (A-I) together with the global length distribution of all new RNA features (J). Antisense RNAs (red) and sense RNAs (grey) relative to the features annotated in GenBank are distinguished.
Remarkably, we could assign up-shifts and related promoter predictions from our tiling array data to 38 of them (Table S6.5). Of these 38 candidate 5’ leader regions, 36 were classified as 5’ regions in our analysis, ncr1889 overlapped the odhA gene, and S1534/ncr1443 classified as “Indep” (independent region with its own up-shift and down-shift). Thus, the 38 of the 40 candidate 5’ leader RNAs can be considered as validated by two independent experimental approaches.

Regulatory 5’ cis-acting structures in B. subtilis

The 5’ cis-acting structures play a key role in transcriptional and post-transcriptional regulation of gene expression in B. subtilis. Generally, they act by sequestering intrinsic terminators or RBSs, located within the 5’ leader region of the target transcript, through conformational changes of the RNA in response to the binding of metabolites, uncharged tRNAs, or proteins.

We collected the known 5’ cis-acting RNAs from different sources. First, in the version 3 of the B. subtilis genome annotation, 57 5’ cis-acting structures are identified by a scan of the genome sequence against the RFAM database of RNA families (36). However, we excluded BSU_misc_RNA_23 (RFAM accession RF00516) because it was categorized as an independent segment (S547) expressed from the opposite strand of its flanking genes. Secondly, 7 cis-acting structures were added from the Genome Reviews version of the genome annotation (37). They were also predicted by a scan of the chromosomal sequence against RFAM. This list was completed by 28 additional features listed in (38) and by guaA (10). In total, we obtained a list of 92 known 5’ cis-acting structures for B. subtilis corresponding to 89 genes (Table S6.6).

In our data set, all the 5’ cis-acting RNAs from the list were found to be significantly expressed, and promoters were mapped for 79 of them. Notably, the 79 promoters are sigA-dependent, with only glmS having alternative promoters (U122.K, U124.L). Interestingly, five genes also have alternative transcription start sites located very near the CDS, suggesting that these genes can be transcribed without the 5’ cis-acting RNA structure: mgtE (U1073.K), ptsG (U1137.GF), metK (U2367.GF), sacX (U3024.A), and yxjA (U3083.GF).

Because of the variety of conditions tested in our study, 58 of the known 5’ cis-acting structures exhibited a clear pattern of transcriptional regulation by a termination/antitermination mechanism (see the example of pbuE, http://www.basysbio.eu/bsubtranscriptome/seb). Down-shifts were detected for 48 of them, and ten more could be added after a visual inspection of the transcription profiles. Remarkably, predicted intrinsic terminators were associated with this type of pattern for all of the 58 features (SOM 6). Furthermore, 13 intrinsic terminators were predicted among the 34 remaining 5’ cis-acting structures, thus pointing at new candidate genes potentially regulated by transcription attenuation (Table S6.6).

(SOM 15) Antisense RNAs in B. subtilis

Origins of asRNAs

In our study, 13% (560/4245) of the protein-coding genes were targeted by asRNAs (Table S11), which is the same proportion that has been reported for Mycoplasma pneumoniae based on the analysis of 43 tiling arrays relating to four time series (growth curve, heat shock, DNA damage, and cell cycle arrest) (2).

The classification of the segments summarized in Table 1 reveals that many antisense transcripts (62%) arise in transcriptional contexts corresponding to incomplete termination of the transcription (categories 3’PT, 3’NT, Indep-NT and Inter). The promoter classification (Table 2, Table S4) was analyzed to determine which categories of promoters are responsible for antisense transcription (Table S14). The prevalence of SigA dependent transcription is
Figure S17: Condition-dependent expression profiles of new RNA features in transcriptional context categories. For each category of RNA feature (described in Table 1), the number of RNA features expressed in the upper 30% (plain line) and expressed above the 5x background (dashed line) is represented across the conditions (x-axis detailed in Figure S4).
much lower for asRNAs than for protein coding genes: only 52% of the asRNAs are predicted to be transcribed from a SigA promoter whereas this fraction is 74% for protein coding genes (Table S14). This trend is most pronounced for the classes of asRNAs that have their own promoters (annotated Indep and Indep-NT in Table1 and Table S5) as only a small minority (23%) is predicted to be SigA-dependent. Overall, 82% (347/423) of the antisense transcripts are accounted for by incomplete termination of transcription or by initiation of transcription from promoters controlled by alternative sigma factors (Table S14). Specifically, these 347 asRNAs comprise the following categories (as listed in Table S14): 53, 5’ AltSig-dep.; 12, 3’UTR AltSig-dep.; 40, 3’ NT; 70, 3’PT; 17, Indep AltSig-dep; 64 Indep-NT ; 89, Inter ; 2, Intra AltSig-dep.

AsRNAs are differentially expressed across conditions

For each class of new mRNA features listed in Table1, the number of expressed (≥ 5 times above background (5x)) and highly expressed (upper 30%) features is represented across the conditions (Figure S17). The 5’, 3’UTR, Inter, and Intra classes (1306/1583), which are embedded in the structure of most TUs, displayed condition-dependent expression profiles similar to that of all genes, and the number of highly expressed features varied only weakly through conditions (Figure S17). In contrast, for the RNA features corresponding to new TUs (Indep, Indep-NT) and to 3’ extended mRNAs (3’PT, 3’NT), the number of highly expressed features varied markedly with conditions (Figure S17). Interestingly, 71% (198/277) of this subgroup (Indep, Indep-NT, 3’PT, 3’NT) were antisense RNAs and represented 47% of all antisense RNAs (Table 1).

Expression above the 5x threshold of asRNAs took place in all conditions, with a median of ~80 asRNA expressed per condition (Figure S18A), and some of these asRNAs were expressed constitutively (Figure S18B). However, the number of highly expressed asRNAs (upper 30%) was strongly dependent upon the conditions, such as competence, sporulation, various stresses, and growth at a high density and on a solid surface (Figure 3). Under these conditions, the total number of expressed segments decreased markedly (Figure S17, “All”), raising the possibility that expression of asRNAs could artifactually become more apparent. However, a general decrease in gene expression was not systematically associated with an increase in the expression of asRNA (for example, germination early stage and growth in D-gluconate). Furthermore, while antisense transcription was particularly prominent during sporulation and in response to stresses, the expression levels of many individual asRNAs decreased (Figure 3B, Figure S18B), indicating that within a condition asRNAs can display a variety of expression profiles. These observations strongly support the differential expression of asRNAs across the conditions.

Interestingly, the maximum and median expression level across conditions for antisense segments tends to be lower than for protein-coding genes, this is true for SigA and non-SigA dependent asRNAs (Figure S19). The expression of the CDSs facing asRNAs is also less likely to reach a very high expression level.

Correlations between expression profiles of sense-antisense pairs

Because of the multiple mechanisms of action of asRNAs (12), an indicated correlation between sense/antisense expression cannot be directly related to a particular mechanism of action of the asRNAs. However, correlations between sense/antisense expressions are relevant to describe the data. For example, transcription interference defined as transcription from one promoter blocks transcription from the second promoter, is expected to result in a negative correlation. However, many other mechanisms by which asRNAs act on their targets (transcription attenuation, modulation of endo-and exonuclease activity, modulation of ribosome binding) can affect positively or negatively the amount of the target RNA (12). Here, when expression levels of asRNAs and cognate sense transcripts are considered across conditions, sense and antisense
Figure S18: Condition-dependent expression profiles of asRNAs compared to the background.
A) Number of antisense transcription segments with expression level above the 5x background threshold. All asRNAs (blue dashed line) and the Indep and Indep-NT RNAs transcribed independently of the annotated features (black dashed line) are represented. The dotted horizontal line represents the average number of asRNAs expressed. The x-axis representing the conditions is as in Figure S4. B) Heat map representation of antisense transcription. The expression of each segment is shown as a horizontal line. The color scale indicates the expression level relative to the background. This figure can be compared with Figure 3.

Figure S19: Relative distributions of expression levels for mRNAs and asRNAs.
The distributions of expression levels for asRNAs, coding mRNAs covered by asRNAs, and all mRNAs are compared. Left panel: maximum expression level across the 269 hybridizations. Right panel: median expression level across the 269 hybridizations. First row: all asRNAs, CDSs faced by asRNAs, and all CDSs. Second row: only SigA-dependent transcripts (asRNAs, CDSs faced by asRNAs, and CDSs). Third row: transcripts dependent on the major alternative sigma factors (SigB, E, F, G, and K). For each distribution, we show a kernel density estimate (lines) and the median (arrows). With respect to the maximum and the median expression level across conditions, expression of asRNAs tends to be lower than expression of CDSs (first row). This trend cannot be explained by the overrepresentation of non SigA-dependent transcripts among asRNAs as it is also observed when SigA-dependent asRNAs and the asRNAs controlled by the major alternative sigma factors are considered separately (second and third rows).
transcripts are more often negatively correlated (47%, p-value cut-off 0.05, Pearson correlation) than positively correlated (30%). Comparison with random pairs of CDSs (36% negative vs. 40% positive correlations) confirms that the excess of negative correlation should be interpreted as the relationship between sense and antisense expression (Figure S20). The excess of negative correlation is due to the frequent configuration where a non-SigA asRNA faces a SigA sense RNA (37% of the sense-antisense pairs). In these cases, 58% negative correlation versus 20% positive correlation are observed, but a small excess of negative correlation still remains detectable when only the other configurations are considered (41% negative vs. 35% positive correlation). Importantly, the observed enrichment in negative correlation between sense and antisense RNAs does not necessarily imply a functional role of asRNAs. It might also result from a selective pressure to avoid potentially deleterious effects mediated by asRNAs. Information on each pair of sense-antisense transcripts can be found in Table S11.

**Evolutionary conservation of the promoters driving antisense transcription**

The asRNAs from the Indep and Indep-NT categories are not associated with previously annotated CDSs and, according to our hypothesis (see main text), they represent the subset of asRNAs most likely to arise from spurious transcription initiation. We investigated whether the promoters of these asRNAs are evolutionarily conserved in a similar fashion as the promoters for sense RNAs. Importantly, our analysis was based solely on nucleotide sequences of related Bacillus species because transcriptome data from these species, which could have been used for a direct comparison of promoter conservation, are presently not available.

We aligned the 3242 *B. subtilis* sequences already used for sigma factor binding site discovery (101 bp around each up-shift, see SOM10) against 17 complete genomes of other Bacillus species (EBI Genome Review Release 130.0) including *Bacillus subtilis* subsp. spizizenii (FN597644_GR, CP001215_GR, CP002207_GR, CP002394_GR, CP001176_GR, CP000764_GR, AP006627_GR, BA000004_GR, AE017333_GR, CP001982_GR, CP001878_GR, CP000813_GR, CP001791_GR, CP002183_GR, CP001903_GR, CP002017_GR, CP000903_GR). Using the FastA alignment algorithm (88) with ktup=3 and other parameters set to default values, the fraction of the query sequences in alignments (E-value ≤ 0.001) varied from 89.3%

![Figure S20](image.png)

**Figure S20: Correlation between the expression levels of sense and antisense RNA pairs.**

Pearson correlation coefficients between expression levels of sense-antisense pairs (kernel density estimate shown in green) are compared to correlation coefficients between random pairs of CDSs (black). For sense-antisense pairs, the contribution of the pairs corresponding to a non-SigA asRNA facing a SigA mRNA (37% of the pairs) is further distinguished from the pairs in other configurations (dashed vs. dotted green line). Correlations that can be considered statistically significant (p-value<0.05) are outside the dashed area. The observed excess of negative correlation is mostly due to non-SigA antisense transcripts facing SigA sense transcripts (dashed green line). However, the other configurations (dotted green line) also tend to display an excess of negative correlation.
to 2.0% while the global average nucleotide divergence in the aligned regions varied between 0.06 and 0.30 differences per bp. Within the aligned regions, the average rates of nucleotide divergence were computed separately for the positions within the predicted sigma factor binding sites and for positions in the surrounding sequences (the 101 bp with -10 and -35 boxes are reported in Table S4). The average rates of nucleotide divergence were determined for our subset of asRNA promoters and for the whole set of promoters (Figure S23A). To capture in a single summary statistic the difference between sigma factor binding sites and their background, a “sigma factor binding site divergence index” was calculated as the ratio between the average rate of nucleotide divergence inside and outside the predicted binding sites (Figure S23B). We find that for moderate evolutionary distances (Bacillus amyloliquefaciens and Bacillus atrophaeus), the sigma factor binding sites associated with Indep and Indep-NT asRNAs (99 up-shifts regions listed in Table S5) are not better conserved than the surrounding sequences. In contrast, when all the promoters are considered, the sigma factor binding sites appear much more conserved than the surrounding sequences. The lack of conservation of asRNAs promoters is statistically significant as revealed by the comparison with a 95% confidence interval for the divergence index under the null hypothesis that the promoter regions responsible for asRNAs are drawn randomly from the whole set of 3242 up-shift regions (Figure S23B). This confidence interval was built by using 1000 randomly sampled sets of size 99 with exactly the same composition as the initial set in terms of the different categories of sigma factors.

Biological relevance of asRNAs

We propose that many asRNAs arise from spurious transcription initiation by alternative sigma factors and/or from imperfect control of transcription termination (see main text).

Figure S23: Evolutionary conservation of promoters responsible for Indep and Indep-NT asRNAs

A) Average rates of nucleotide divergence inside (plain line) and outside (dashed line) sigma factor binding sites computed for the 101 bp promoter regions driving transcription of Indep and Indep-NT asRNAs (green) and for the whole promoter dataset (black). The different Bacillus species are positioned along the x-axis according to the fraction of nucleotides that could be aligned unambiguously (E-value≤0.001). B) The sigma factor binding site divergence index is the ratio between the rates of nucleotide divergence inside and outside sigma factor binding sites (SOM 15). This index is represented for the promoters of the Indep and Indep-NT asRNAs (plain green line) and for the whole promoter dataset (plain black line). The median (dashed black line) and a 95% confidence interval (grey area) are also indicated for the divergence index under the null hypothesis that the promoters generating Indep and Indep-NT asRNAs are similar to the whole set of promoters.
• With respect to initiation, the recognition by the sigma-bound RNAP of the degenerate sequence motifs that constitute promoters is modulated via secondary DNA structures and protein factors (39). Evolutionary constraints are expected to have shaped the genome sequence to minimize spurious transcription initiation, which could generate asRNAs with potentially deleterious effects. The use of many alternative sigma factors inevitably multiplies the number of sequences potentially recognized as promoters. In this context, spurious transcription initiated by the constitutive housekeeping SigA might be less prevalent than that initiated by alternate sigma factors, which have more sporadic activities. This hypothesis of a lower evolutionary burden of spurious transcription initiated by alternative sigma factors would explain our observation that asRNAs are often initiated by alternative sigma factors. Of note, this does not imply that alternative sigma factors are intrinsically more error-prone than SigA. Furthermore, the finding that sigma factor binding sites in the promoter regions of Indep and Indep-NT asRNAs tend to be less conserved than those in the promoter regions of protein-coding genes (see above paragraph) further supports our hypothesis.

• With respect to termination, we identified many asRNAs arising from incomplete or missing termination sites, suggesting imperfect control of transcription termination (SOM 7). We also found that Rho plays a key role in limiting asRNAs by preventing transcription beyond the boundaries of a subset of TUs.

Such mechanisms to generate asRNAs are compatible with the observed differential expression of asRNAs (Figure 3). In addition, the observation that asRNAs tend to be expressed in lower amounts than protein-coding RNAs (Figure S19) suggests that some of them may not have a regulatory function. Indeed, previous studies on plasmids and transposons have revealed that asRNAs must be produced in excess amounts over the target RNAs to effectively act as regulators (40, 41, 42). We found in our dataset two asRNAs (S885, S975) targeting annotated genes, which were classified as never expressed under any of the conditions tested (see the B. subtilis expression data browser at http://www.basysbio.eu/bsubtranscriptome/seb). S885 is expressed from a SigK-dependent promoter and S975 is expressed from SigB- and SigEF-dependent promoters (Table S11). These are unambiguous examples of situations where antisense transcription takes place independently of a regulation mediated by the pairing of the asRNA with the cis-encoded target RNA, which is in keeping with the hypothesis of spurious initiation of asRNAs.

Notably, our hypothesis does not rule out the possibility that some asRNAs have functional roles. Indeed, B. subtilis contains at least one asRNA-regulated toxin-antitoxin system (type I toxin-antitoxin system) comprising a toxic peptide (TxpA) and the RNA antitoxin (RatA/S976) (35). In addition, two other toxin genes of B. subtilis were identified as antisense targets: the gene encoding NdoA (Table S6.4) which is inhibited by the antitoxin NdoAI (43), and the gene encoding SdpC whose toxic effects are neutralized by the immunity protein SdpI (44). In both cases, the expression of the asRNA (S163 and S1294, respectively) showed negative correlation with the respective mRNA. In addition, the potential antisense targets of B. subtilis include genes involved in a wide range of biological functions of which 26 genes were previously shown to be essential (9). Many of the potential targets are known to be regulated at the transcriptional level by non-antisense mechanisms indicating that one function of antisense regulation might be to fine-tune the levels of certain proteins as, for example, in the cases of the virulence protein MgtC of Salmonella enterica (45) and the urease B subunit of Helicobacter pylori (46). In B. subtilis, the ilvBHC-leuABCD operon encoding enzymes for branched-chain amino acid biosynthesis is subject to exceptionally complex regulation (47). This includes a post-transcriptional mechanism, i.e. differential stability of individual mRNA species resulting from mRNA processing (48). In agreement with the studies by Rasmussen et al. (7) and Irnov et al. (10) (SOM5, Table S6.4), we identified an asRNA (S1070) that is complementary to the ilvH, ilvC and leuA genes, potentially being one example of the modulation of the levels of extensively regulated targets. Our study also
revealed a remarkably high number of 24 transcriptional regulators including two sigma factors (SigA and SigL) targeted by asRNAs. Interestingly, genes encoding transcription factors are also known to be regulated by asRNAs in other bacteria (49, 50, 51, 52).
Chapter 2

References


