Deposited research article

Identification of attenuation and antitermination regulation in prokaryotes

Warren C. Lathe III*†‡, Mikita Suyama*†, and Peer Bork†‡

Addresses: †EMBL, Meyerhofstr. 1, D69012 Heidelberg, Germany. ‡Max Delbrück Center for Molecular Medicine, D13092 Berlin-Buch, Germany.

*These authors contributed equally.

Correspondence: Warren C. Lathe III. E-mail: lathe@embl-heidelberg.de

Posted: 30 April 2002

Genome ***Biology*** 2002, **3(6)**:preprint0003.1-0003.60

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/6/preprint/0003

Received: 24 April 2002

This is the first version of this article to be made available publicly. This article was submitted to *Genome* ***Biology*** for peer review.

Identification of attenuation and antitermination regulation in prokaryotes

by Warren C. Lathe III*†‡, Mikita Suyama*†, and Peer Bork†‡

April 18, 2002

*These authors contributed equally

†EMBL, Meyerhofstr. 1, D69012 Heidelberg, Germany.

‡Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany.

Abstract

Many operons of biochemical pathways in bacterial genomes are regulated by processes called attenuation and antitermination. Though the specific mechanism can be quite different, attenuation and antitermination in these operons have in common the termination of transcription by a RNA 'terminator' fold upstream of the first gene in the operon. In the past, detecting regulation by attenuation or antitermination has often been a long process of experimental trial and error, on a case by case basis. We report here the prediction of over 290 upstream regions of genes with attenuation or antitermination regulation structures in the completed genomes of *Bacillis subtilis* and *Escherichia coli* for which extensive experimental studies have been done on attenuation and antitermination regulation. These predictions are based on a computational method devised from characteristics of known terminator fold candidates and benchmark regions of entire genomes. We extend this methodology to 24 additional complete genomes and are thus able to give a more complete picture of attenuation and antitermination regulation in bacteria.

Background

The control of gene expression can occur at many points in the transcription and translation of the genes of bacterial operons. Two mechanisms of operon regulation of great interest are "attenuation" and "antitermination"[1-5]. These mechanisms regulate the early termination of transcription of a wide variety of operons in diverse species. Classically, attenuation occurs when the transcribed RNA upstream of an operon has the ability to fold into two mutually-exclusive RNA-fold structures, one which is termed an antiterminator and the other a terminator. If the terminator hairpin loop is allowed to fold,

transcription is ultimately halted. Alternatively, if the antiterminator structure folds, the terminator is precluded from folding and transcription of the operon proceeds. The mechanisms that alternate between these two RNA folds (terminators and antiterminators) are quite diverse. Regulation by antitermination (not to be confused with the alternative antiterminator fold of attenuation) can be differentiated from attenuation by the fact that alteration of the transcription complex (rather than alternate RNA structures) decreases the efficiency of downstream terminators. Though, in reality, the boundary between these two types of regulation is not distinct[3].

Attenuation and antitermination mechanisms have both been described in a wide variety of regulatory and biochemical pathways. These include operons involved in aminoacyl-tRNA biosynthesis, catabolic metabolism, amino-acid biosynthesis, ABC transport systems, ribosomal structural peptides and several others. They have been characterized in genomes as disparate as the low-GC gram-positive *Bacillis subtilis* and the proteobacteria *Escherichia coli.* The precise mechanisms that cause the attenuation or antitermination of these operons can be quite distinct. For example, the *trp* operons of *E. coli*[4,5] and *B. subtilis*[6,7], though both regulated by attenuation, are controlled by quite different mechanisms. Other operons, such as the structural ribosomal S10 operon of *E. coli*[8,9] are regulated by yet a different mechanism. Even between closely related species, the attenuation and antitermination, and upstream regulatory sequences can be entirely different.

Yet, one common and necessary feature of most experimentally described attenuation and antitermination mechanisms is an intrinsic terminator RNA fold structure[2,3]. The stem-

loop structure of an intrinsic terminator has been well described[10] and the understanding of the mechanisms of termination has made great progress in recent years[11,12]. This structure is not only found at the location of 'standard' termination of transcription at the end of transcriptional units, but is also, by definition a part of attenuation and antitermination regulation. The major characteristics of this standard terminator structure is that it is relatively short, is energetically stable, has a G/C rich stem, contains a small terminal loop structure and, importantly, also contains a run of U residues on the 3' side of the stem-loop structure[10]. These characteristics of intrinsic terminators have been used in the past to predict terminator structures at the end of transcriptional units (operons) and thus assist in the prediction of transcriptional units in complete genomes[10,13] and on a limited basis for predicting regulatory attenuators[14]. Here we focus on using the characteristics and position of intrinsic terminators to predict and characterize attenuation and antitermination regulation in operons of *B. subtilis* and *E. coli.* These mechanisms of regulation are well described in these two genomes. We extend this characterization to an additional 24 genomes representative of the diversity of eubacteria and archeabacteria to give a broader picture of attenuation and antitermination regulation in prokaryotes and in a more automated and extensive manner than previously achieved.

Results

Characterization of attenuators in *B. subtilis* and *E. coli*

An extensive literature search for operons in *B. subtilis* regulated by attenuation or antitermination was conducted and 46 such operons were found. These range from the experimentally well described *trp* operon to those operons where terminator structures have been found and attenuation is expected though not well characterized experimentally[15,16] (for a full list see http://www.bork.embl-heidelberg.de/Docu/attenuation). These 46 known terminator structures were employed to determine common characteristics of *B. subtilis* attenuation terminators. Using these characteristics, we screened upstream regions of 3650 *B. subtilis* genes (using procedures described in Materials and Methods) for terminator folds. Forty-three of the original 46 known terminators found in the literature search were retained in this screening. An additional 1117 upstream folds that fit our criteria were also obtained. In addition, as a control, we used the same filtering and folding methodology on intergenic regions after the sequences were shuffled randomly (952 folds of randomly shuffled sequences were obtained after filtering).

The resulting folds of all intergenic regions and shuffled sequences obtained after filtering were plotted in terms of their stability and length (Figure 1). The known terminator folds lie in a cluster clearly separate and distinct from those folds of randomly shuffled sequences. Terminator folds are of a lesser free energy (ΔG) in relation to length than predicted folds of random sequences. A similar pattern of two easily separated clusters emerges when comparing known terminator structures with folded intragenic regions in which terminator are not expected to be found (data not shown).

Using principal component analysis, we determined the greatest variance of the randomly shuffled sequences. This can give us a measure (using standard deviation) of which folds are significantly different from folds of random sequences (see Materials and Methods). Of the 1160 folds, a total of 203 folds of intergenic regions obtained in our screen fall below the 2nd deviation line ($Z \leq -2$) derived from the principal component. These are thus considered significantly different from random folds and possible terminations sites of attenuation or antitermination regulation. Forty-two of these are the known attenuation terminators folds (of the original 43 known folds maintained after filtering). Thus we are able to obtain 91.3% (42/46) of the known and experimentally characterized attenuation and antitermination sites using our filter and significance measure. Additionally, the filter and significance measure screens out over 97.7% (930 of 952) of the folds of random sequences. One hundred and sixty-one (203 total excluding 42 known) folds under the line ($Z \leq -2$) are folds not yet analyzed experimentally and could be predicted to be attenuation terminator structures.

A detailed investigation found many of these predictions are strongly supported as a putative attenuation or antitermination sites by genomic context such as the presence of putative promoter sequences, upstream location of putative and known operons, etc. Two terminator structures upstream genes *ydbJ* and *yqhI* serve as detailed examples of how genomic context can inform and strongly support the predictions made in Table 1 (Figure 2). Gene *ydbJ* of *B. subtilis* is listed as hypothetical with homology to an ABC transporter gene (ATP-binding protein involved in copper transport). The gene immediately downstream, *ydbK*, has homology to membrane spanning permeases. Using STRING (a search tool for find recurring instances of neighboring genes[17]), orthologs

of these two genes are also found in the same order in transcriptional units of 15 other distantly related genomes, suggesting the possibility these genes form an operon. These genes appear to be in a typical ABC transporter operon configuration and several ABC transporter operons are known to be regulated by attenuation in *B. subtilis*[15,16]. The *ydbJ* upstream region also has a putative promoter sequence and predicted folds using RNAfold (See Materials and Methods) of the entire upstream sequence suggest it can fold in complex possible antitermination folds (data not shown). Based on this context, we predict this is an ABC transporter operon regulated by attenuation. The second example, *yqhI*, is the first gene of a run of three genes all having homology to glycine biosynthesis genes in a putative transcriptional unit. This run of three genes also has orthologs found as neighbors in other genomes[17]. Many amino acid biosynthesis operons in *B. subtilis* are known to be regulated by attenuation[16], thus supporting this prediction.

In order to see if the observed patterns hold for the only other genome in which attenuation or antitermination is well studied and experimentally described, we also applied the same methodology to upstream regions of genes in the *E. coli* genome for which 16 operons have been described as being regulated by attenuation or antitermination. As can be seen in Figure 3, the known *E. coli* attenuation and antitermination terminator structures have similar properties as those of *B. subtilis.* 15 of the 16 known attenuators were maintained after filtering. The significance measure separates 14 of these *E. coli* terminators from random folds as seen in Figure 3. As in *B. subtilis,* using the ($Z \leq -2$) line as a measure of significance, we are able to predict attenuation for 146 regions (Figure 3 and Table 2).

Extension of analysis to 26 genomes

Analysis of *B. subtilis* and *E. coli* suggest that a broader survey of bacterial genomes might prove useful in both the prediction of attenuation and antitermination regulation in these genomes and the characterization of the evolution and distribution of these mechanisms of regulation. Twenty-four completed genomes were selected for this survey based on their broad distribution across the evolutionary spectrum (Table 3). The intergenic regions of each of these genomes were analyzed using the same methods and filters as with *B. subtilis* and *E. coli* and predicted attenuation and antitermination terminator folds similarly obtained.

As shown in Table 3, there is a wide distribution of the number of putative attenuation and antitermination regulatory sites in the surveyed genomes. These range from 5 in *Mycobacterium tuberculosis* to 275 in *Clostridium acetobutylicum* (Table 3). Earlier attempts to predict standard transcription termination sites at the end of transcription units give similar results. Interestingly, the results for standard transcription terminators correlate with ours. As was found in Ermolaeva *et. al*[13] with standard terminators at the end of transcription units (this paper studied terminators at end of ORFs and did not target upstream regions, thus filtering out possible attenuators), some of the highest number of occurrences of attenuation and antitermination sites in our survey are similarly found in the genomes of *E. coli, H. influenze, D. radiodurans* and *B. subtilis* and the lowest number of occurrences in such genomes as *H. pylori*, and *M. tuberculosis* (genomes reported in their survey).

At first glance, this would seem to suggest that many genomes do not use the same mechanisms of termination for the standard transcription termination and do not use attenuation or antitermination in regulation. This is likely the case in some genomes. Yet, if the number of upstream intergenic regions is plotted against the number of predicted sites, a strong positive correlation is shown (Figure 4). The smaller the number of genes and intergenic regions a genome has, the lower the occurrence of predicted terminators (both standard transcription terminators and attenuation/antitermination regulatory terminators). This indicates that the low numbers of both standard termination and regulatory termination in many genomes is due to a much reduced genome size and the reduction of the number of regulatory operons, and not necessarily to the reliance on different mechanisms of termination and regulation.

There is a clear outlier with a much lower than expected number of putative terminators seen in Figure 4, *Mycobacterium tuberculosis*. This genome has a much lower occurrence of putative attenuation and antitermination sites than would be suggested by its size and the number of intergenic regions. A recent paper by Unniraman *et al*.[18] concludes that *M. tuberculosis* uses a different mechanism of termination that utilizes terminator structures without the poly-U tail necessary in other genomes. Thus the reduced number of poly-U containing terminator structures in relation to the number of intergenic regions can be explained by *M. tuberculosis'* reliance on a different mechanism of termination. This does not necessarily prove there is no attenuation or antitermination type regulation in *M. tuberculosis*. However, it does indicate that either the loss of the standard mechanism of termination in this genome has reduced if not eliminated attenuation or

antitermination in *M. tuberculosis* or alternatively, an attenuation-like mechanism could exist in this genome that utilizes the *M. tuberculosis* non-standard terminator.

All other of the 25 genomes surveyed have putative attenuation or antitermination regulation sites. Even the lowest number of predicted attenuation or antitermination sites found in *M. genitalium* are a significant proportion of possible regulatory intergenic regions, the low number is easily accounted for by this genome's relatively small size and few intergenic regions and transcriptional units. These results suggest that attenuation and antitermination regulation is a possibly ubiquitous mechanism of regulation in prokaryotes with few exceptions.

Genome Size and Attenuation

If the GC content of a genome is compared with the number of predicted attenuators based on randomly shuffled sequence, GC content does somewhat correlate with the number of predicted attenuators, which would be expected since a poly-U run is required in the filters. In Figure 5a, folds from randomly shuffled intergenic sequences of our 26 genomes were plotted by the number of filtered folds per intergenic region in relation to number of intergenic regions. If the number of filtered folds was completely random, there should be a relatively constant number of sites per region in relation to the number of regions. As seen in figure 5a, this is not completely the case. The number of filtered folds per region obtained from randomly shuffled sequences is dependent on the GC content of the genome. Low-GC content genomes have a slightly higher per region number of folds than do genomes of around 50% GC content and high-GC content genomes have much lower number than both. This is expected from random sequences

filtered for stem-loop structures containing poly-U runs.

Even when taking into account the GC content of *M. tuberculosis,* it has a reduced number of predicted attenuators in relation to the other high-GC genomes (Figure 5b). In fact, Figure 5b (predicted attenuators of actual intergenic sequences) shows that the strongest determinate of the number of predicted attenuators per intergenic region is not GC content but rather genome size (more specifically the number of intergenic regions). In general, not only do larger genomes have a greater absolute number of predicted attenuators, but have a greater occurrence of predicted attenuators per region. If GC content is equal in two genomes, the larger genome is more likely to have a higher number of predicted attenuators per intergenic region. Previous reports have suggested similar phenomena in regulatory proteins, large genomes appear to have a larger proportion of their total number of genes that code for proteins which contain regulatory motifs[19]. Interestingly, discounting the archaebacteria and high GC content genomes, a genome of about 1500 intergenic regions appears to be the threshold at where the frequency of regulatory attenuators increases in a genome.

Distribution and Conservation of Attenuators in Gram positive Bacteria

Seven genomes of gram-positive bacterias (*B. subtilis, B. halodurans, L. innocua, S. aureus, C. acetobutylicum, L. lactis, and S. pneumoniae*) were analyzed to see whether the attenuation terminators are conserved in front of the orthologs. The number of predicted attenuation terminators for the genes known to be regulated in *B. subtilis* and their orthologs in the other six genomes are listed in Table 4a. The genomes are sorted by phylogenetic distance from *B. subtilis* calculated by amino acid sequences of the shared

orthologs among these genomes. The closest one to the *B. subtilis* is *B. halodurans* and the averaged number of amino acid substitutions per site is 0.238, and the most distant one is *S. pneumoniae* and the averaged number of amino acid substitutions per site is 0.422. For the 42 genes listed in Table 4a, the numbers of orthologs that are found in the other genomes vary little from genome to genome: The highest and the lowest numbers of orthologs are 31 in *L. lactis* and 26 in *S. aureus* and *C. acetobutylicum*, respectively. This is mainly because these 42 genes carry some basic functions such as aminoacyl-tRNA synthesis. On the other hand, the numbers of predicted attenuation termination structures vary significantly: In *B. halodurans,* 22 orthologous genes have predicted attenuation termination structures, while only 4 orthologous genes have the predicted structures in *S. pneumoniae*. This indicates that the absence or presence of regulation by attenuation is much more weakly conserved than the gene or operons presence.

The same trend holds true for the predicted attenuation termination structures other than known ones (Table 4b). There are 105 orthologous gene groups that have at least one other genome containing a predicted attenuator structure upstream an orthologous gene. Restricting to the orthologs that have predicted attenuators in *B. subtilis* (35 groups), the highest and the lowest numbers of shared orthologs of genes known to be regulated by attenuation or antitermination in *B. subtilis* are 28 (*L. innocua*) and 18 (*S. pneumoniae*), respectively. The numbers of predicted attenuation termination structures, however, vary more. While there are 13 genes with predicted structures in *B. halodurans*, which is the closest species to *B. subtilis* among the six gram-positive bacterias, only 2 genes have predicted structures in *S. pneumoniae*.

Although there is weak conservation of attenuators as a whole, predicted attenuation termination structures and the order of their downstream genes are conserved for some groups of genes. One of such example is *infC-rpmI-rplT* operon (figure 6a). No attenuation termination structure is predicted in the upstream region of *infC* in *S. pneumoniae* (Table 4b). Closer look at this region by BLAST[20] revealed that the N-

terminal of *infC* is over predicted in 27 bases. By adding the 27 bases to the intergenic region in the upstream, we found a stable stem-loop structure that followed by poly-U residues also in *S. pneumoniae* (Figure 6b). Even in this example however, there are considerable differences among species in the relative position of the stem-loop structures and sequence conservation. Moreover, even between the phylogenetically closest pair, *B. subtilis* and *B. halodurans*, the distances from the end of the stem to the start codon of *infC* are 69 and 37 bases, respectively, and only the common segments found in the stem are GUGUGGGN{*x*}CCCACAC (*x* = 12 in *B. subtilis* and *x* = 9 in *B. halodurans*). Among all the seven genomes, there is only a weak similarity, GYGGG (GACGG in *C. acetobutylicum*) in the stem region.

Conservation of predicted attenuation termination structures is also observed in the upstream regions of the possible operon containing *nusA* gene (Figure 7a). Four out of seven genomes contain predicted attenuator structures in upstream of the hypothetical protein (*ylxS* in *B. subtilis*). Stem-loop structures are also found in the rest of three genomes, although these structures do not pass the filters. The location of the structures to the transcription start site of the downstream gene and sequences themselves vary significantly in this example also. In these stem sequences, the segment GUGGG (GAGCG in *L. lactis* and GAGGC in *S. pneumoniae*) is conserved in the predicted operon containing *nusA* gene (Figure 7b). Interestingly, the 5-base segments are identical or very similar to the segments in the stem-loop structures located in the upstream of *infC* (figure 6b). The proteins encoded the genes in these two operon are involved in transcription. The conservation of the sequence segments in the predicted attenuation terminator structures for *infC-rpmI-rplT* operon and the operon containing *nusA* implies that there exists a common regulatory mechanism that recognizes the stem-loop structure and this would regulate both operons in the same manner.

Distribution and Conservation of Attenuators in Proteobacteria

Several aspects of the conservation of attenuators are immediately apparent from our analysis of gram-positive bacteria . First, the distribution of attenuation or antitermination regulation is not well conserved across gram-postive baceria and additionally, even in conserved regulatory systems, sequence and structure conservation is weak. The same holds true for proteobacteria. Of the 14 genes in *E. coli* (see Table 5a) known to be regulated by attenuation or antitermination, none have attenuators predicted upstream orthologs in all of the four other proteobacteria genomes. Six have attenuators predicted upstream orthologs in at least one of the other four genomes. Three are genes that have orthologs in all four other genomes, but these have no predicted attenuators. The remaining five genes in *E. coli* have either no known orthologs in the other genome or orthologs have a spotty distribution and no predicted attenuators. Closer inspection by hand confirms this conclusion. Table 5b is a list of all predicted attenuators in each of the five genomes of the gamma division of proteobacteria in which a similar attenuator is predicted for an ortholog of another genome. As shown in this table, attenuation and antitermination appears to be poorly conserved as a mechanism of regulation in analogous operons in proteobacterial genomes. Of the total of 475 genes and their orthologs in these five genomes that have predicted attenuators, only 36 are shared upstream orthologs of two or more genomes (Tables 3, 5a and 5b).

Previous research concerning specific systems have reported that attenuation and antitermination regulation in some operons in *E. coli* are only mildly conserved across gamma division proteobacteria. The regulation *rpsJ* operon [21] and the *trpE* and *pheA* operons[22] of *E. coli* have been shown to have a spotty distribution and weakly

conserved across proteobacteria. As shown in Tables 2 and 5, we have been able to extensively extend this analysis of attenuation and antitermination to most such systems in proteobacteria, and have shown that this holds true for all known attenuation and antitermination regulatory mechanisms in *E. coli* and other predicted mechanisms in additional gamma division genomes. An example is given in figure 8 of the low sequence conservation of attenuators and regulation. In figure 8a, one of the more conserved attenuators is shown for that of the *hisG* operon. This operon and regulatory mechanism is well characterized in *E. coli*[23] and our analysis predicts similar mechanisms of attenuation regulation in *V. cholerae* and *H. influenzae*. The predicted attenuators have conserved position (at approximately 40-50bp upstream start codon of *hisG* gene), and stem sequence. Though the surrounding intergenic regions are not possible to align, *V. cholerae* and *H. influenzae* do have possible amino acid leader sequences with a run of histidines that is characteristic of the attenuation regulation mechanism in *E. coli*. Predicted attenuators were not found in the other three gamma subdivision probacteria genomes of *P. aeruginosa*, *N. meningitidus* and *X. fastidiosa*. In *P. aeruginosa* the intergenic region upstream of the *hisG* ortholog is only 17bp in length, in *X. fastidious* the orthologous gene overlaps with the ORF upstream, and though the analogous *N. meningitidus* intergenic region is of sufficient length, no attenuator is predicted.

Discussion

In summary, attenuation terminators reveal a striking pattern distinct from both folds of randomly shuffled sequences and intragenic regions. In relation to their length, terminator folds have a much lower free energy (ΔG) than random folds or those within cistronic

regions. This enables us to differentiate and predict many novel attenuation regulation sites in a variety of putative operons and would be a 5-fold increase in the number of known attenuation structures in *B. subtilis* and *E. coli*. This measure works in two highly divergent species with distinct mechanisms of attenuation and antitermination, and different GC content. Hence, it is feasible to extend such analysis to all bacterial genomes. Extending the study to a diverse collection an additional 24 complete genomes suggests that attenuation and antitermination is likely used in most genomes, with the possible exception of *M. tuberculosis,* as a form of regulation.

The standard transcription termination mechanism likely came early in the evolution of bacteria and regulation by attenuation and antitermination most probably arose by co-opting existing terminators and the transcription termination mechanism. How and when attenuation and antitermination has evolved in individual genomes and taxa and how this mechanism arose in specific operons and biochemical systems is a question that can now be further analyzed and is a subject of future work.

This study also allows us to make strong predictions for specific instances of attenuation and antitermination regulation. Previously, Merino et al.[14] published a chapter in TITLE looking at orthologous genes of genes known to be regulated by attenuation and antitermination and found a significant number of putative attenuators. These were reported also on a web site (**http://cmgm.stanford.edu/~merino**). The results of our research reported here confirm most of those predictions. In addition, as shown in this paper, since the conservation of attenuation and antitermination regulation across a taxa is weak, looking at orthologous genes in other genomes will miss many potential

attenuators. This report greatly extends the predictions of attenuation and antitermination. These predictions can be very useful in directing future research. As proof of point, one such prediction made by our study was for attenuation regulation in the gene *ctrA* (*pyrG*) in *B. subtilis.* In the course of our study a recent report confirmed this prediction[23].

This research also enables a better understanding of the evolution and distribution of attenuation and antitermination regulation. Such predictions can be very beneficial in directing research into operon regulation, assisting in predicting gene function, understanding the evolution of regulation in general and heightening our understanding of regulons.

Materials and Methods

Genome sequence data

Genome sequences and their annotations were obtained from GenBank[24] (species and accession numbers: *A. fulgidus* AE000782; *B. burgdorferi* AE000783; *B. halodurans* BA000004; *B. subtilis* AL009126; *Buchnera* sp. AP000398; *C. acetobutylicum* AE001437; *C. jejuni* AL111168; *C. pneumoniae* AE001363; *D. radiodurans* chromosome 1 AE000513; *E. coli* K-12 U00096; *H. influenzae* L42023; *H. pylori* J99 AE001439; *L. innocua* AL592022; *L. lactis* AE005176; *M. genitalium* L43967; *M. jannaschii* L77117; *M. tuberculosis* AL123456; *N. meningitidis* MC58 AE002098; *P. abyssi* AL096836; *P. aeruginosa* AE004091; *S. aureus* Mu50 BA000017; *S. pneumoniae*

AE005672; *Synechocystis* sp. AB001339; *T. maritima* AE000512; *V. cholerae* chromosome 1 AE003852; *X. fastidiosa* AE003849).

RNA folding and filters

For each gene in a genome, we collected upstream sequence segments up to 300 residues or the the neighboring ORF. ORFs with less than 50 amino acids were considered as intergenic regions since some attenuation mechanisms are coupled with the synthesis of leader peptides. The total number of these segments was 3560 in *B. subtilis* and 3613 in *E. coli*. Stem-loop structures in these segments were predicted by using the RNAfold program[25]. As a reference, each upstream sequence was shuffled to produce random sequences with the same base composition and folded in the same manner. Using the characteristics of the known structures, we derived three filters based on location and poly-U runs which were applied to the collected folds to optimize the possibility of finding new attenuation structures (Figure 9). The filtering process retained 44 of the 46 known attenuation terminators. The three filters used, based on known attenuation terminators in *B. subtilis* and *E. coli* are: (i) Poly-U stretches (≥4 Us) must be located within from the 10 residues at the top of the stem to 15 residues downstream of the stem: (ii) the length of the 3' sequence from the end of the stem to the start position of the downstream gene must be ≤170 and (iii) the length of the 5' sequence from the beginning of the intergenic region to the 5' start of the stem must be ≥30, if the upstream gene is in the same orientation as the downstream (to partially eliminate 'standard' transcription terminator structures). Available public programs for transcription termination prediction, such as TransTerm[13] were not useful in this analysis. The program takes the direction of neighboring genes into account and distance filters, and could not be applied to the

prediction of termination structures located upstream of a gene.

To futher support the prediction of attenuation terminator, we also applied promoter prediction: if a promoter is predicted in the upstream of a predicted attenuation terminator, then it is more likely to be the real attenuation terminator. NNPP version 2.2[26, 27] was used for the prediction of prokaryotic promoters.

Significance measurement

To evaluate the significance of stem-loop structures in upstream sequences, we used the distribution of those structures in the randomly shuffled sequences, which has the same base composition and only the order of the bases are randomized. First, all the stem-loop structures found in the shuffled sequences are plotted according the their stability and stem length (figures 1 and 3). Then the line running along the largest variance is calculated by principal component analysis[28]. Using the standard deviation which is calculated from distribution of stem-loop structures in the shuffled sequence around the line, *Z*-score is calculated for each stem-loop structure. We took those stem-loop structures in the upstream sequences with $Z \leq -2$ as significant structures.

Identification of orthologous genes

To identify orthologous gene pair among a pair of genomes, first we carried out all-against-all comparison between the sets of proteins, each of which is from a whole genome. We used the BLASTP program[20] for this comparison. Only the hits with BLAST $E \leq 0.001$ are collected as significant hits. Then, amang those significant hits, a

pair of genes are defined as ortholog if the pair satisfies the "bi-directional best hit"[17]. For a group of more than 2 genomes, a group of genes, each of which is from a genome, are defined as ortholog if all possible pairs of genes satisfy the bi-directional significant best hit.

References

1. Henkin TM **Control of transcription termination in prokaryotes**. *Annu Rev Genet* 1996, **30**:35-57.

2. Yanofsky C: **Transcription attenuation: once viewed as a novel regulatory strategy.** *J Bacteriol* 2000, **182**:1-8.

3. Wagner R: *Transcription Regulation in Prokaryotes.* Oxford Oxford University Press, 2000.

4. Yanofsky C: **Attenuation in the control of expression of bacterial genomes**. *Nature*1981, **289**:751-758.

5. Yanofsky C, Konan KV, Sarsero JP: **Some novel transcription attenuation mechanisms used by bacteria.** *Biochimie* 1996, **78**:1017-1024.

6. Babitzke P: **Regulation of tryptophan biosynthesis: Trp-ing the TRAP or how *Bacillus subtilis* reinvented the wheel**. *Mol Microbiol* 1997, **26**:1-9.

7. Du H, Yakhnin A, Dharmaraj S, Babitzke P: ***trp* RNA-binding attenuation protein-5' stem-loop RNA interaction is required for proper transcription attenuation control of the *Bacillus subtilis trpEDCFBA* operon.** *J Bacteriol* 2000, **182**:1819-1827.

8. Allen T, Shen P, Samsel L, Liu R, Lindahl L, Zengel JM: **Phylogenetic analysis of L4-mediated autogenous control of the S10 ribosomal protein operon.** *J Bacteriol* 1999, **181**:6124-32.

9. Zengel JM, Lindahl L: **A hairpin structure upstream of the terminator hairpin required for ribosomal protein L4-mediated attenuation control of the S10 operon of *Escherichia coli.*** *J Bacteriol* 1996, **178**:2383-238.

10. Carafa YdA, Brody E, Thermes C: **Prediction of rho-independent *Escherichia coli* transcription terminators: A statistical analysis of their RNA stem-loop structures**. *J Mol Biol* 1990, **216**:835-858.

11. Wilson KS, von Hippel PH: **Transcription termination at intrinsic terminators: The role of the RNA hairpin.** *Proc Natl Acad Sci USA* 1995, **92**:8793-8797.

12. Yarnell WS, Roberts JW: **Mechanism of intrinsic transcription termination and antitermination.** *Science* 1999, **284**:611-615.

13. Ermolaeva MD, Khalak HG, White O, Smith HO, Salzberg SL: **Prediction of transcription terminators in bacterial genomes.** *J Mol Biol* 2000, **301**:27-33.

14. Merino E, Yanofsky C: **Regulation by Termination-Antitermination: a Genomic Approach**, *in Bacillis subtilis and its closest relatives: From Genes to Cells.* Washington D.C.: American Society of Microbiology, 2001.

15. Chopin A, Biaudet V, Ehrlich SD: **Analysis of the Bacillus subtilis genome sequence reveals nine new T-box leaders.** *Mol Microbiol* 1998, **29**:661-669.

16. Grundy FJ, Henkin TM: **The S box regulon: a new global transcription termination control system for methionine and cystein biosynthesis genes in gram-positive bacteria.** *Mol Microbiol* 1998, **30**:737-749.

17. Snel B, Lehmann G, Bork P, Huynen M: **STRING: a web-server to retrieve and display the repeatedly occurring neighborhood of a gene**. *Nucleic Acids Res* 2000*,* **28**:3442-3444.

18. Unniraman S, Prakash R, Nagaraja V: **Alternate paradigm for intrinsic transcription termination in eubacteria.** *J Biol Chem* 2001, **276**:41850-41855.

19. Stover CK et. al: **Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen.** *Nature* 2000, **406**:959-64.

20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25:**339-3402.

21. Allen T, Shen P, Samsel L, Liu R, Lindahl L, Zengel J.M: **Phylogenetic analysis of L4-mediated autogenous control of the S10 ribosomal protein operon**. *J Bacteriol* 1999, **181**:6124-32.

22. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS: **Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria.** *J Mol Microbiol Biotechnol* 2001, **3**:529-43.

23. Meng Q, Switzer R: **Regulation of Transcription of the Bacillus subtilis *pyrG* Gene, encoding cytidine triphosphate synthetase.** *J Bacteriol* 2001, **183:** 5513-5522.

24. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2002, **30:**17-20.

25. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P: **Fast folding and comparison of RNA secondary structures.** *Monatsh Chem* 1994, **125:**167-188.

26. Reese, MG: **Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome.** *Comput Chem* 2001, **26:**51-56.

27. Neural Network Promoter Prediction:
[http://www.fruitfly.org/seq_tools/promoter.html]

28. Afifi AA, Clark V: Computer-aided multivariate analysis. 3rd ed: Baca Raton, Florida: Chapman & Hall, 1996.

Figure Legends

Figure 1

Stability and length distributions of stem-loop structures in upstream sequence segments of *B. subtilis.*

The red line shows the largest variance (see Materials and Methods) derived from stem-loop structures in shuffled sequences. Light blue lines give the significance measurements based on standard deviation. The definition for each point together with the orientation of neighboring genes are shown in upper right panel.

Figure 2.

Schematic drawing of the neighborhood and predicted structures for the *B. subtilis* genes *ydbJ* and *yqhI*.

Genes are signified by colored arrows and are in orientation of transcription in relation to orientation of reference gene (*ydbJ* or *yqhI*). Large blue stem-loop cartoons signify predicted terminator fold in attenuation, 't' is an annotated standard terminator fold. Intergenic regions are drawn to scale and bp lengths of these are given underneath figure.

Figure 3.

Stability and length distributions of stem-loop structures in upstream sequence segments in *E. coli.*

The red line shows the largest variance (see Materials and Methods) derived from stem-loop structures in shuffled sequences. Light blue lines give the significance measurements based on standard deviation. The definition for each point together with the orientation of neighboring genes are shown in upper right panel.

Figure 4.

Graph of the number of intergenic regions vs. the number of putative attenuation and antitermination sites in all 26 genomes surveyed. Several genomes with known attenuation or antitermination are labeled for comparison as is *M. tuberculosis* and the Archaea. The dashed line is a exponential trendline.

Figure 5: Genome Size and Regulation.
(a) Intergenic sequences of 26 genomes were randomly shuffled, folded and filtered using reported method to obtain putative 'attenuators'. The number of these shuffled and filtered folds per intergenic region were plotted for each genome against the number of intergenic regions. The correlation, if random, should remain constant and independent of genome size. Blue spheres represent proteobacteria and Bacillis species in our survey, beige are archaeabacteria and green the rest. Spheres are in size in proportion to the genome's GC content and GC content is labeled within each sphere. The number of random folds per intergenic region is a function of GC content as would be expected from filtering for folds with poly-U runs. Genomes with known attenuation or antitermination are labeled as is the genome known not to use attenuators with poly-U runs in termination. (b) Intergenic sequences of 22 genomes were folded and filtered for possible attenuators and indication of attenuation or antitermination regulation. The number of these predicted attenuators per intergenic region is compared to the number of intergenic regions in the genome. In contrast to folds of randomly shuffled sequences, the strongest determinate for the frequency of attenuation is genome size (number of intergenic regions and genome size are strongly correlated). Colors and labeling are the same as in 5a.

Figure 6.

Predicted attenuation termination structure in upstream region of putative *infC-rpmI-rplT* operon.

(a) Order of genes. Only intergenic regions are drawn to scale and the length of intergenic regions are given below the line. Orthologous genes are indicated in the same colors. Hypothetical genes and the other non-orthologous genes are indicated by "hyp" and their gene IDs, respectively. Abbreviation for genomes: *Bs*, *B. subtilis*; *Bh*, *B. halodurans*; *Li*, *Listeria innocua*; *Sa*, *Staphylococcus aureus*; *Ca*, *Clostridium acetobutylicum*; *Ll*, *Lactococcus lactis*; *Sp*, *Streptococcus pneumoniae*. (b) Predicted attenuation termination structures. Base pairs are indicated by red dots between the base codes. Base numbering shows the distance from the start codon of the down stream gene. Poly-Us just down stream of the stem-loop structure is colored in green. Weakly conserved segments are colored in red. Abbreviation for genomes is the same as in (a).

Figure 7.

Predicted attenuation termination structure in upstream region of *ylxS* gene.

(a) Order of genes. Predicted stem-loop structures with statistical significance are indicated in blue, and the other structures that neither pass the filters nor have less significance are indicated in red. For the other explanation, see legend to figure 6a. (b) Predicted attenuation termination structures. See legend to figure 6b for the explanation.

Figure 8.

Predicted attenuation termination structure in upstream region of *HisG* gene in *E. coli*.

(a) Order of genes. Predicted stem-loop structures with statistical significance are indicated in blue. For the other explanation, see legend to figure 6a. Abbreviations for genomes: Ec*, Escherichia coli;* Hi*, Haemophilus influenzae;* Vc*, Vibrio cholerae;* Pa*,*

Pseudomonas aeruginosa; Xf*, Xylella fastidiosa;* Nm*, Neisseria meningitidis*. (b) Predicted attenuation termination structures. See legend to figure 6b for the explanation.

Figure 9.

Schematic drawing of the analysis of upstream sequence segments and definition of filters as described in Materials and Methods.

Tables

Table 1.

ID	Z-score	Known	Upstream promoter[a]	Gene[b]
BS0929	-8.895	√		glycerol-3-phosphate dehydrogenase (*glpD*)
BS2825	-8.473	√	√	acetolactate synthase (acetohydroxy-acid synthase) (large subunit) (*ilvB*)
BS1310	-6.836			similar to chaperonin (*ykkC*)
BS0983	-6.733		√	hypothetical protein (*yhaW*)
BS3920	-6.468	√	√	PTS beta-glucoside-specific enzyme IIABC component (*bglP*)
BS1257	-6.332			RNA polymerase PBSX sigma factor-like (*xpf*)
BS2733	-6.238	√		alanyl-tRNA synthetase (*alaS*)
BS0013	-6.126	√	√	seryl-tRNA synthetase (*serS*)
BS2184	-5.942		√	hypothetical protein (*yphP*)
BS3966	-5.840		√	hypothetical protein (*iolD*)
BS2749	-5.839	√		histidyl-tRNA synthetase (*hisS*)
BS2520	-5.815	√	√	glycyl-tRNA synthetase (alpha subunit) (*glyQ*)
BS3396	-5.767	√	√	similar to amino acid permease (*yvbW*)
BS0215	-5.526	√	√	glycerol-3-phosphate permease (*glpT*)
BS2255	-5.344			hypothetical protein (*ypiA*)
BS2600	-5.342			similar to phage-related protein (*yqbK*)
BS1544	-5.239	√	√	isoleucyl-tRNA synthetase (*ileS*)
BS3710	-5.196		√	CTP synthetase (*ctrA*)
BS3895	-5.191		√	similar to pyrimidine nucleoside transport protein (*yxjA*)
BS3798	-5.164	√	√	PTS sucrose-specific enzyme IIBC component (*sacP*)
BS2614	-5.152		√	hypothetical protein (*yqaQ*)
BS2484	-5.148		√	similar to 5-formyltetrahydrofolate cyclo-ligase (*yqgN*)
BS1319	-5.124	√	√	cobalamin-independent methionine synthase (*metC*)
BS2682	-4.990		√	hypothetical protein (*yraL*)
BS3750	-4.958	√	√	threonyl-tRNA synthetase (*thrZ*)
BS0183	-4.938		√	NADH dehydrogenase (subunit 5) (*ndhF*)
BS3440	-4.890	√	√	levansucrase (*sacB*)
BS2715	-4.841		√	similar to formate dehydrogenase (*yrhE*)
BS2204	-4.819	√	√	xanthine phosphoribosyltransferase (*xpt*)
BS2139	-4.756	√	√	similar to *N*-acetylmuramoyl-L-alanine amidase (*yomC*)
BS2803	-4.750	√	√	valyl-tRNA synthetase (*valS*)

BS0093	-4.747	√	√	serine acetyltransferase (*cysE*)
BS3877	-4.696		√	hypothetical protein (*yxkD*)
BS1357	-4.680	√	√	hypothetical protein (*ykrT*)
BS0178	-4.675		√	L-glutamine-D-fructose-6-phosphate amidotransferase (*glmS*)
BS1658	-4.644		√	DNA polymerase III (alpha subunit) (*polC*)
BS3894	-4.630		√	hypothetical protein (*yxjB*)
tRNA-Arg	-4.609			tRNA-Arg (*trnJ-Arg*)
BS0535	-4.505			similar to antibiotic resistance protein (*ydfB*)
BS3839	-4.472	√		tyrosyl-tRNA synthetase (*tyrZ*)
BS2961	-4.430	√	√	tyrosyl-tRNA synthetase (*tyrS*)
BS0441	-4.399			hypothetical protein (*ydbB*)
BS1143	-4.382	√		tryptophanyl-tRNA synthetase (*trpS*)
BS1737	-4.375		√	similar to ribonucleoprotein (*ymaA*)
BS0254	-4.309	√	√	hypothetical protein (*yczA*)
BS1971	-4.297		√	similar to adenosylmethionine-8-amino-7-oxononanoate aminotransferase (*yodT*)
BS0643	-4.279	√	√	phosphoribosylaminoimidazole carboxylase I (*purE*)
BS3026	-4.270	√	√	leucyl-tRNA synthetase (*leuS*)
BS2889	-4.249	√		threonyl-tRNA synthetase (*thrS*)
BS3690	-4.191		√	hypothetical protein (*ywlC*)
BS1188	-4.176	√	√	similar to cystathionine gamma-synthase (*yjcI*)
BS0521	-4.158		√	hypothetical protein (*ydeI*)
BS2452	-4.143		√	similar to aminomethyltransferase (*yqhI*)
BS3604	-4.128		√	hypothetical protein (*ywrE*)
BS1313	-4.086	√		gamma-glutamyl kinase (*proB*)
BS3269	-3.989	√	√	similar to ABC transporter (ATP-binding protein) (*yusC*)
BS0423	-3.974		√	hypothetical protein (*ydaH*)
BS2504	-3.896		√	similar to transcriptional regulator (Fur family) (*yqfV*)
BS3900	-3.884	√	√	endo-beta-1,3-1,4 glucanase (*bglS*)
BS0038	-3.883	√	√	methionyl-tRNA synthetase (*metS*)
BS2319	-3.879		√	hypothetical protein (*ypuF*)
BS1683	-3.845			similar to phage-related protein (*ymfE*)
BS1881	-3.805		√	phosphoenolpyruvate synthase (*pps*)
BS3513	-3.805		√	putative membrane protein (*csbA*)
BS0432	-3.796			hypothetical protein (*ydaO*)
BS3355	-3.795			hypothetical protein (*yvaI*)
BS2858	-3.784	√	√	phenylalanyl-tRNA synthetase (alpha subunit) (*pheS*)
BS3668	-3.780		√	hypothetical protein (*ywmD*)
BS0927	-3.776	√		glycerol uptake facilitator (*glpF*)
BS0281	-3.739		√	hypothetical protein (*ycdC*)

BS2306	-3.714			RNA polymerase ECF-type sigma factor (sigma-X) (*sigX*)
BS0143	-3.709		√	RNA polymerase (alpha subunit) (*rpoA*)
BS0449	-3.687		√	similar to ABC transporter (ATP-binding protein) (*ydbJ*)
BS1094	-3.668		√	hypothetical protein (*yitC*)
BS1370	-3.645		√	motility protein A (*motA*)
BS3785	-3.618		√	hypothetical protein (*ywdL*)
BS0104	-3.612		√	ribosomal protein L10 (BL5) (*rplJ*)
BS3313	-3.584			similar to iron-binding protein (*yvrC*)
BS0716	-3.572		√	hypothetical protein (*yetH*)
BS1889	-3.524		√	response regulator aspartate phosphatase (*rapK*)
BS0673	-3.492			hypothetical protein (*yerQ*)
BS0002	-3.480			DNA polymerase III (beta subunit) (*dnaN*)
BS0880	-3.473		√	transcriptional regulator (*senS*)
BS1548	-3.473	√	√	transcriptional attenuator and uracil phosphoribosyltransferase activity (minor) (*pyrR*)
BS1409	-3.467		√	hypothetical protein (*ykuH*)
BS2881	-3.464			initiation factor IF-3 (*infC*)
BS1651	-3.412			uridylate kinase (*smbA*)
BS3888	-3.412	√	√	hypothetical protein (*yxjH*)
BS3889	-3.412	√	√	hypothetical protein (*yxjG*)
BS0925	-3.407		√	similar to adenosylmethionine-8-amino-7-oxononanoate aminotransferase (*yhxA*)
BS1205	-3.351			hypothetical protein (*yjdG*)
BS3784	-3.335			hypothetical protein (*spsA*)
BS3093	-3.325		√	hypothetical protein (*yuaJ*)
BS1470	-3.265		√	hypothetical protein (*yktD*)
BS1549	-3.265	√		uracil permease (*pyrP*)
BS2841	-3.262		√	aspartokinase II (*lysC*)
BS2903	-3.222			DNA polymerase I (*polA*)
BS3446	-3.202		√	hypothetical protein (*yvdQ*)
BS2984	-3.202			hypothetical protein (*ytmQ*)
BS1920	-3.201		√	similar to ATP-dependent DNA helicase (*yocI*)
BS3343	-3.192		√	hypothetical protein (*yvgV*)
BS3731	-3.192		√	hypothetical protein (*ywiA*)
BS3278	-3.188			similar to 3-hydroxyacyl-CoA dehydrogenase (*yusL*)
BS3412	-3.180			transcriptional regulator (LacI family) (*lacR*)
BS2278	-3.156			hypothetical protein (*yphE*)
BS3914	-3.141		√	hypothetical protein (*yxiF*)
BS1550	-3.120	√	√	aspartate carbamoyltransferase (*pyrB*)
BS0878	-3.120		√	thiamin biosynthesis protein (*thiA*)
BS2775	-3.111		√	similar to sodium/proton-dependent alanine carrier protein (*yrbD*)
BS1673	-3.111			dipicolinate synthase subunit A (*spoVFA*)

BS2569	-3.086		√	site-specific DNA recombinase (*spoIVCA*)
BS2713	-3.075		√	similar to formate dehydrogenase (*yrhG*)
BS2539	-3.075			heat-shock protein (*dnaJ*)
BS3719	-3.068			similar to cardiolipin synthetase (*ywiE*)
BS2711	-3.061		√	similar to methyltransferase (*yrhH*)
BS1166	-3.059	√		transcriptional regulator (*tenA*)
BS1446	-3.057			aminopeptidase (*ampS*)
BS0503	-3.033		√	hypothetical protein (*yddM*)
BS1455	-3.014		√	hypothetical protein (*ykzG*)
BS3164	-2.969		√	transcriptional regulator (*comQ*)
BS1360	-2.966	√	√	similar to ribulose-bisphosphate carboxylase (*ykrW*)
BS1536	-2.942		√	similar to acetylornithine deacetylase (*ylmB*)
BS3133	-2.940		√	similar to polyribonucleotide nucleotidyltransferase (*yugI*)
BS3635	-2.930		√	flagellar basal-body rod protein (*flhO*)
BS1659	-2.908		√	hypothetical protein (*ylxS*)
BS2216	-2.896			hypothetical protein (*ypsA*)
BS0115	-2.894		√	ribosomal protein S10 (BS13) (*rpsJ*)
BS0637	-2.878		√	hypothetical protein (*yebB*)
BS2627	-2.867			similar to transcriptional regulator (phage-related) (Xre family) (*yqaE*)
BS0745	-2.856		√	similar to quinone oxidoreductase (*yfmJ*)
BS1785	-2.852			transcriptional regulator (*lexA*)
BS1737	-2.805		√	similar to ribonucleoprotein (*ymaA*)
BS1525	-2.800		√	cell-division initiation protein (*divIB*)
BS2138	-2.782			hypothetical protein (*yomD*)
BS2162	-2.776		√	hypothetical protein (*yokC*)
BS1335	-2.740		√	similar to transcriptional regulator (MarR family) (*ykoM*)
BS0241	-2.658		√	similar to histidine permease (*ybgF*)
BS2715	-2.649		√	similar to formate dehydrogenase (*yrhE*)
BS1855	-2.649	√	√	similar to phosphoglycerate dehydrogenase (*yoaD*)
BS2986	-2.643		√	hypothetical protein (*ytmP*)
BS1321	-2.642			hypothetical protein (*ispU*)
BS0386	-2.640		√	similar to transcriptional regulator (TetR/AcrR family) (*ycnC*)
BS2425	-2.616			similar to exodeoxyribonuclease VII (large subunit) (*yqiB*)
BS0057	-2.589			similar to amino acid transporter (*yabM*)
BS3957	-2.586			similar to ABC transporter (ATP-binding protein) (*yxdL*)
BS1003	-2.585		√	Hit-like protein (*hit*)
BS3350	-2.580			hypothetical protein (*yvaC*)
BS1659	-2.573		√	hypothetical protein (*ylxS*)
BS4050	-2.567		√	hypothetical protein (*yybP*)
BS2617	-2.559		√	hypothetical protein (*yqaN*)

BS0806	-2.555	√	acetoin dehydrogenase E1 component (TPP-dependent alpha subunit) (*acoA*)
BS0636	-2.553	√	GMP synthetase (*guaA*)
BS1871	-2.546	√	hypothetical protein (*yoaS*)
BS2152	-2.537	√	hypothetical protein (*yolA*)
BS1490	-2.535	√	cytochrome caa3 oxidase (subunit II) (*ctaC*)
BS3210	-2.518		transcriptional regulator (*paiA*)
BS0558	-2.516	√	hypothetical protein (*ydgC*)
BS1585	-2.516		similar to phosphoglycerate dehydrogenase (*yloW*)
BS3625	-2.513		transcriptional regulator (DeoR family) (*glcR*)
BS2512	-2.483	√	hypothetical protein (*yqfN*)
BS0164	-2.480		similar to transcriptional regulator (AraC/XylS family) (*ybbB*)
BS2895	-2.471	√	hypothetical protein (*ytcF*)
BS2632	-2.456	√	hypothetical protein (*yrkS*)
BS0804	-2.447	√	hypothetical protein (*yfjM*)
BS0632	-2.404	√	similar to cation efflux system membrane protein (*yeaB*)
BS2632	-2.392	√	hypotheitcal protein (*yrkS*)
BS2713	-2.387	√	similar to formate dehydrogenase (*yrhG*)
BS0561	-2.386	√	ATP-binding transport protein (*expZ*)
BS1373	-2.372	√	hypothetical protein (*ykvJ*)
BS0888	-2.371	√	hypothetical protein (*ygaO*)
BS2239	-2.369		ketopantoate hydroxymethyltransferase (*panB*)
BS2746	-2.356	√	hypothetical protein (*yrvN*)
BS3568	-2.347		UDP-glucose:polyglycerol phosphate glucosyltransferase (*tagE*)
BS4087	-2.342	√	similar to formate dehydrogenase (*yyaE*)
BS0745	-2.309	√	similar to quinone oxidoreductase (*yfmJ*)
BS3473	-2.308	√	similar to mutator MutT protein (*yvcI*)
BS2596	-2.308	√	similar to phage-related protein (*yqbN*)
BS1319	-2.296	√	cobalamin-independent methionine synthase (*metC*)
BS1472	-2.294	√	hypothetical protein (*ylaA*)
BS3265	-2.287	√	similar to ABC transporter (ATP-binding protein) (*yurY*)
BS3666	-2.250		required for formate dehydrogenase activity (*narQ*)
BS2225	-2.236	√	hypothetical protein (*yppD*)
BS3000	-2.232	√	hypothetical protein (*ytfP*)
BS1181	-2.230	√	hypothetical protein (*yjcB*)
BS1735	-2.220		hypothetical protein (*ymzC*)
BS0668	-2.210	√	hypothetical protein (*yerL*)
BS1065	-2.204		similar to DNA exonuclease (*yirY*)
BS3199	-2.201	√	hypothetical protein (*yuiF*)
BS1302	-2.182		hypothetical protein (*ykgB*)
BS2921	-2.181		hypothetical protein (*ytoI*)

BS3142	-2.171		√	similar to multidrug-efflux transporter (*yuxJ*)
BS2729	-2.163		√	similar to caffeoyl-CoA *O*-methyltransferase (*yrrM*)
BS0313	-2.157			hypothetical protein (*ycgI*)
BS3278	-2.157			similar to 3-hydroxyacyl-CoA dehydrogenase (*yusL*)
BS1386	-2.151		√	similar to heavy metal-transporting ATPase (*ykvW*)
BS0245	-2.139		√	similar to two-component sensor histidine kinase [YcbB] (*ycbA*)
BS3196	-2.124		√	hypothetical protein (*yuiI*)
BS0826	-2.115		√	similar to metabolite transport protein (*yfiG*)
BS1536	-2.115		√	similar to acetylornithine deacetylase (*ylmB*)
BS1142	-2.104			hypothetical protein (*yjbA*)
BS2376	-2.057	√	√	similar to pyrroline-5-carboxylate reductase (*yqjO*)
BS2920	-2.018			hypothetical protein (*ytpI*)
BS1878	-2.018			beta-lactamase precursor (*penP*)

Table 1. Predicted attenuators in the genome of *B. subtilis*

First column is gene ID, second is Z score based on method in Materials and methods. Third column is checked if an attenuator has already been shown experimentally. Fourth column indicates if a promoter was predicted from NNPP program. Fifth column lists gene description.

a The NNPP program (cutoff=0.8) was used for the upstream promoter prediction

b GenBank annotation was used for gene definitions and gene names

Table 2.

ID	Z-score	Known	Upstream promoter[a]	Gene[b]
b4119	-9.279			alpha-galactosidase (*melA*)
b2019	-7.560	√	√	ATP phosphoribosyltransferase (*hisG*)
b3828	-7.143			regulator for *metE* and *metH* (*metR*)
b3543	-7.009			dipeptide transport system permease protein 1 (*dppB*)
b4118	-6.911			regulator of melibiose operon (*melR*)
b2425	-6.168		√	thiosulfate binding protein (*cysP*)
b0560	-5.838		√	bacteriophage DNA packaging protein (*nohB*)
b4385	-5.755			hypothetical protein (*yjjJ*)
b3866	-5.703			hypothetical protein (*yihI*)
b0689	-5.697			putative pectinase (*ybfP*)
b1548	-5.638		√	homolog of Qin prophage packaging protein NU1 (*nohA*)
b1825	-5.573		√	hypothetical protein
b0611	-5.528			RNase I, cleaves phosphodiester bond between any two nucleotides (*rna*)
b2622	-5.470			prophage CP4-57 integrase (*intA*)
b3066	-5.002			DNA biosynthesis; DNA primase (*dnaG*)
b0654	-4.844			glutamate/aspartate transport system permease (*gltJ*)
b1978	-4.772		√	putative factor
b3767	-4.529	√		acetolactate synthase II, large subunit, cryptic, interrupted (*ilvG_1*)
b2119	-4.518		√	hypothetical protein (*yehL*)
b2792	-4.444			hypothetical protein
b0902	-4.443		√	pyruvate formate lyase activating enzyme 1 (*pflA*)
b3871	-4.352		√	putative GTP-binding factor (*yihK*)
b2937	-4.347		√	agmatinase (*speB*)
b0939	-4.300			putative chaperone (*ycbR*)
b1431	-4.280		√	hypothetical protein
b0890	-4.275			cell division protein (*ftsK*)
b2632	-4.177		√	putative GTP-binding protein (*yfjP*)
b0048	-4.168		√	dihydrofolate reductase type I (*folA*)
b0336	-4.133			cytosine permease/transport (*codB*)
b2517	-4.076		√	hypothetical protein (*yfgB*)
b3438	-4.018			regulator of gluconate (*gnt*) operon (*gntR*)
b2599	-3.945	√	√	chorismate mutase-P and prephenate dehydratase (*pheA*)
b3671	-3.923	√	√	acetolactate synthase I, valine-sensitive, large subunit (*ilvB*)
b2775	-3.907		√	putative transport protein (*yqcE*)
b2609	-3.891		√	30S ribosomal subunit protein S16 (*rpsP*)
b3181	-3.889			transcription elongation factor: cleaves 3' nucleotide of paused mRNA (*greA*)

b3983	-3.861		√	50S ribosomal subunit protein L11 (*rplK*)
b3722	-3.823	√		PTS beta-glucosides enzyme II, cryptic (*bglF*)
b0002	-3.788	√	√	aspartokinase I, homoserine dehydrogenase I (*thrA*)
b0196	-3.759			regulator in colanic acid synthesis; interacts with RcsB (*rcsF*)
b0007	-3.711			inner membrane transport protein (*yaaJ*)
b3513	-3.694			putative membrane protein (*yhiU*)
b3088	-3.669		√	putative transport protein (*ygjT*)
b4178	-3.630		√	hypothetical protein (*yjeB*)
b3752	-3.573	√		ribokinase (*rbsK*)
b2643	-3.572			hypothetical protein (*yfjX*)
b0074	-3.559	√	√	2-isopropylmalate synthase (*leuA*)
b2096	-3.530		√	tagatose-bisphosphate aldolase 1 (*gatY*)
b0170	-3.506	√	√	elongation factor EF-Ts (*tsf*)
b0610	-3.506		√	regulator of nucleoside diphosphate kinase (*rnk*)
b3813	-3.503	√		DNA-dependent ATPase I and helicase II (*uvrD*)
b1851	-3.446			6-phosphogluconate dehydratase (*edd*)
b2359	-3.395		√	hypothetical protein
b4245	-3.387	√		aspartate carbamoyltransferase, catalytic subunit (*pyrB*)
b4377	-3.374		√	hypothetical protein (*yjjU*)
b3298	-3.362		√	30S ribosomal subunit protein S13 (*rpsM*)
b0606	-3.358			alkyl hydroperoxide reductase, F52a subunit (*ahpF*)
b0018	-3.350		√	Gef protein interferes with membrane function when in excess (*gef*)
b2958	-3.329			hypothetical protein (*yggN*)
b3822	-3.303			ATP-dependent DNA helicase (*recQ*)
b0241	-3.292			outer membrane pore protein E (*phoE*)
b0680	-3.260			glutamine tRNA synthetase (*glnS*)
b0404	-3.233			putative glycoprotein (*yajB*)
b3337	-3.232		√	hypothetical protein (*yheA*)
b3190	-3.181			hypothetical protein (*yrbA*)
b0860	-3.147			arginine 3rd transport system periplasmic binding protein (*artJ*)
b1922	-3.115		√	flagellar biosynthesis; regulation of flagellar operons (*fliA*)
b2836	-3.103			2-acyl-glycerophospho-ethanolamine acyltransferase; acyl-acyl-carrier protein synthetase (*aas*)
b3915	-3.087			putative transport system permease protein (*yiiP*)
b3310	-3.083		√	50S ribosomal subunit protein L14 (*rplN*)
b0119	-3.018			hypothetical protein (*yacL*)
b3748	-3.005			D-ribose high-affinity transport system; membrane-associated protein (*rbsD*)
b2313	-2.990		√	membrane protein required for colicin V production (*cvpA*)
b1352	-2.967		√	hypothetical protein (*ydaD*)
b0441	-2.961			putative protease maturation protein (*ybaU*)

b3161	-2.947			tryptophan-specific transport protein (*mtr*)
b1614	-2.888		√	hypothetical protein (*ydgA*)
b3528	-2.862		√	uptake of C4-dicarboxylic acids (*dctA*)
b2514	-2.812		√	histidine tRNA synthetase (*hisS*)
b3390	-2.810		√	shikimate kinase I (*aroK*)
b1253	-2.808			hypothetical protein (*yciA*)
b2724	-2.808		√	probable small subunit of hydrogenase-3, iron-sulfur protein (part of formate hydrogenlyase (FHL) complex) (*hycB*)
b4242	-2.806			Mg2+ transport ATPase, P-type 1 (*mgtA*)
b0208	-2.804			putative transcriptional regulator LYSR-type (*yafC*)
b1479	-2.790		√	NAD-linked malate dehydrogenase (malic enzyme) (*sfcA*)
b1838	-2.749		√	protein phosphatase 1 modulates phosphoproteins, signals protein misfolding (*pphA*)
b0189	-2.741		√	hypothetical protein (*yaeO*)
b1264	-2.717	√		anthranilate synthase component I (*trpE*)
b4313	-2.706		√	recombinase involved in phase variation; regulator for fimA (*fimE*)
b3573	-2.691			hypothetical protein (*yiaI*)
b3642	-2.675	√		orotate phosphoribosyltransferase (*pyrE*)
b0034	-2.646		√	transcriptional regulator of *cai* operon (*caiF*)
b2176	-2.635			hypothetical protein (*rtn*)
b1769	-2.618			putative transport protein (*ydjE*)
b1561	-2.605			hypothetical protein (*rem*)
b1050	-2.605			hypothetical protein (*yceK*)
b2643	-2.604			hypothetical protein (*yfjX*)
b0248	-2.604			hypothetical protein (*yafX*)
b4278	-2.591		√	IS4 hypothetical protein (*yi41*)
b3560	-2.583			glycine tRNA synthetase, alpha subunit (*glyQ*)
b3723	-2.565	√	√	positive regulation of *bgl* operon (*bglG*)
b1689	-2.560		√	hypothetical protein
b3632	-2.528		√	lipopolysaccharide core biosynthesis (*rfaQ*)
b1404	-2.521		√	IS30 transposase (*tra8_2*)
b2014	-2.519		√	putative amino acid/amine transport protein (*yeeF*)
b2924	-2.487			putative transport protein (*yggB*)
b0070	-2.433		√	putative transport protein (*yabM*)
b4037	-2.409		√	periplasmic protein of *mal* regulon (*malM*)
b2359	-2.405			hypothetical protein
b3571	-2.393		√	alpha-amylase (*malS*)
b4082	-2.370			putative membrane protein (*yjcR*)
b0174	-2.360		√	hypothetical protein (*yaeS*)
b4392	-2.359			soluble lytic murein transglycosylase (*slt*)
b2938	-2.347		√	biosynthetic arginine decarboxylase (*speA*)

b2443	-2.331	√	hypothetical protein
b4090	-2.296	√	ribose 5-phosphate isomerase B (*rpiB*)
b2140	-2.289	√	putative regulator protein (*yohI*)
b1861	-2.274	√	Holliday junction helicase subunit B (*ruvA*)
b3196	-2.274	√	hypothetical protein (*yrbG*)
b3054	-2.268		hypothetical protein (*ygiF*)
b1761	-2.252		NADP-specific glutamate dehydrogenase (*gdhA*)
b0763	-2.248	√	molybdate-binding periplasmic protein; permease (*modA*)
b0909	-2.234		putative heat shock protein (*ycaL*)
b3954	-2.233	√	putative ARAC-type regulatory protein (*yijO*)
b3440	-2.232	√	putative regulator (*yhhX*)
b3074	-2.225		putative tRNA synthetase (*ygjH*)
b0605	-2.218	√	alkyl hydroperoxide reductase, C22 subunit; detoxification of hydroperoxides (*ahpC*)
b1053	-2.218		putative transport protein (*yceE*)
b1284	-2.210	√	putative DEOR-type transcriptional regulator
b0068	-2.206		thiamin-binding periplasmic protein (*tbpA*)
b0681	-2.204		hypothetical protein (*ybfM*)
b2150	-2.158	√	galactose-binding transport protein; receptor for galactose taxis (*mglB*)
b3349	-2.153	√	FKBP-type peptidyl-prolyl *cis-trans* isomerase (rotamase) (*slyD*)
b2298	-2.148		putative *S*-transferase (*yfcC*)
b0554	-2.133	√	hypothetical protein (*ybcR*)
b1778	-2.121		hypothetical protein (*yeaA*)
b1829	-2.092	√	heat shock protein, integral membrane protein (*htpX*)
b2685	-2.084		multidrug resistance secretion protein (*emrA*)
b2456	-2.069		detox protein (*cchB*)
b4284	-2.067	√	IS30 transposase (*tra8_3*)
b1002	-2.064	√	periplasmic glucose-1-phosphatase (*agp*)
b0854	-2.059	√	periplasmic putrescine-binding protein; permease protein (*potF*)
b1368	-2.051		putative alpha helix protein
b3088	-2.048	√	putative transport protein (*ygjT*)
b0347	-2.019	√	3-(3-hydroxyphenyl)propionate hydroxylase (*mhpA*)
b1663	-2.018	√	putative transport protein (*ydhE*)

Table 2. Predicted attenuators in the genome of *E. coli*

First column is gene ID, second is Z score based on method in Materials and methods. Third column is checked if an attenuator has already been shown experimentally. Fourth column indicates if a promoter was predicted from NNPP program. Fifth column lists gene description.

a The NNPP program (cutoff=0.8)26,27 was used for the upstream promoter prediction

b GenBank annotation was used for gene definitions and gene names

Table 3.

Genome	No. of intergenic regions	No. of predicted attenuators
Archaea		
Archaeoglobus fulgidus	1684	11
Methanococcus jannaschii	1515	45
Pyrococcus abyssi	1376	12
Eubacteria		
Chlamydiales		
Chlamydia pneumoniae	924	30
Cyanobacteria		
Synechocystis sp.	2957	113
Gram-positive		
Bacillus halodurans	3527	208
Bacillus subtilis	3650	203
Clostridium acetobutylicum	3386	275
Lactococcus lactis	1946	193
Listeria innocua	2565	154
Mycobacterium tuberculosis	3117	5
Mycoplasma genitalium	354	9
Staphylococcus aureus Mu50	2338	180
Streptococcus pneumoniae	1680	147
Proteobacteria (beta subdivision)		
Neisseria meningitidis MC58	1868	118
Proteobacteria (gamma subdivision)		
Buchnera sp.	550	18
Escherichia coli	3613	146
Haemophilus influenzae	1482	107
Pseudomonas aeruginosa	4756	78
Vibrio cholerae	2203	115
Xylella fastidiosa	2116	29

Proteobacteria (epsilon subdivision)		
Campylobacter jejuni	1028	28
Helicobacter pylori J99	1150	29
Spirochaetales		
Borrelia burgdorferi	638	14
Thermas/Deinococcus		
Deinococcus radiodurans	2055	35
Thermotogales		
Thermotoga maritima	1173	23

Table 3. List of all 26 genomes surveyed in this study

Table 4a.

B. subtilis		*B. halodurans*		*L. innocua*		*S. aureus*		*C. acetobutylicum*		*L. lactis*		*S. pneumoniae*		
ID	Z-score	ID[a]	No. of predictions[b]	ID[a]	No. of predictions[b]	ID[a]	No. of predictions[b]	ID[a]	No. of predictions[b]	ID[a]	No. of predictions[b]	ID[a]	No. of predictions[b]	
	No. of predictions[b]	Genec												
BS0929	-8.895	BH1095	1	lin1331	1	SAV1302	0	-	-	L0013	0	SP2185	0	glycerol-3-phosphate dehydrogenase (*glpD*)
BS2825	-8.473	BH3061	1	lin2091	0	SAV2054	0	CAC3169	0	L0078	0	SP0445	0	acetolactate synthase (large subunit) (*ilvB*)
BS3920	-6.468	BH0296	1	lin0026	1	-	-	CAC1407	1	L90678	0	SP0577	0	PTS beta-glucoside-specific enzyme II ABC component (*bglP*)
BS2733	-6.238	BH1267	1	lin1539	1	SAV1618	1	CAC1678	1	L0343	1	SP1383	0	alanyl-tRNA synthetase (*alaS*)
BS0013	-6.126	BH0024	1	lin2890	1	SAV0009	1	CAC0021	0	L150515	1	SP0411	1	seryl-tRNA synthetase (*serS*)
BS2749	-5.839	BH1251	2	lin1555	1	SAV1631	1	CAC2740	0	L0342	0	SP2121	0	histidyl-tRNA synthetase (*hisS*)
BS2520	-5.815	BH1370	1	lin1496	1	-	-	-	-	L101560	0	SP1475	0	glycyl-tRNA synthetase (alpha subunit) (*glyQ*)
BS3396	-5.767	-	-	-	-	-	-	-	-	L18622	1	-	-	amino acid permease (*yvbW*)
BS0215	-5.526	-	-	-	-	SAV0337	0	-	-	L148346	0	-	-	glycerol-3-phosphate permease (*glpT*)
BS1544	-5.239	BH2545	1	lin2127	1	SAV1193	1	CAC3038	1	L0350	1	SP1659	1	isoleucyl-tRNA synthetase (*ileS*)
BS3798	-5.164	BH1856	0	-	-	-	-	CAC0423	1	-	-	-	-	PTS sucrose-specific enzyme II BC component (*sacP*)
BS1319	-5.124	BH0438	1	lin1789	1	SAV0356	0	-	-	L0100	0	SP0585	0	cobalamin-independent methionine synthase (*metC*)
BS3750	-4.958	-	-	-	-	-	-	CAC2362	1	-	-	-	-	threonyl-tRNA synthetase (*thrZ*)
BS3440	-4.890	-	-	-	-	-	-	CAC1772	0	-	-	-	-	levansucrase (*sacB*)
BS2204	-4.819	BH1514	1	lin1998	1	SAV0388	1	CAC0873	0	L159396	1	SP1847	0	xanthine phosphoribosyltransferase (*xpt*)
BS2139	-4.756	-	-	-	-	-	-	-	-	-	-	-	-	*N*-acetylmuramoyl-L-alanine amidase (*yomC*)
BS2803	-4.750	BH3038	1	lin1587	1	SAV1663	1	CAC2399	2	L0351	1	SP0568	0	valyl-tRNA synthetase (*valS*)
BS0093	-4.747	BH0110	1	lin0270	0	SAV0529	1	CAC0687	1	L0087	1	SP0589	0	serine acetyltransferase (*cysE*)
BS1357	-4.680	-	-	-	-	-	-	-	-	-	-	-	-	hypothetical protein (*ykrT*)
BS3839	-4.472	BH3228	1	-	-	-	-	CAC0780	1	-	-	-	-	tyrosyl-tRNA synthetase (*tyrZ*)
BS2961	-4.430	-	-	lin1639	1	SAV1729	1	CAC0637	0	L0359	0	SP2100	0	tyrosyl-tRNA synthetase (*tyrS*)
BS1143	-4.382	BH2870	0	lin2301	1	SAV0996	0	CAC0626	1	L0358	1	SP2229	0	tryptophanyl-tRNA synthetase (*trpS*)
BS0254	-4.309	-	-	-	-	-	-	-	-	-	-	-	-	hypothetical protein (*yczA*)
BS0643	-4.279	BH0623	1	lin1887	0	SAV1064	0	CAC1390	0	L152487	0	SP0053	0	phosphoribosylaminoimidazole carboxylase I (*purE*)
BS3026	-4.270	BH3281	1	lin1769	1	SAV1760	0	CAC0646	1	L0352	0	SP0254	0	leucyl-tRNA synthetase (*leuS*)
BS2889	-4.249	BH3141	1	lin1594	1	SAV1683	1	-	-	L0357	1	SP1631	0	threonyl-tRNA synthetase (*thrS*)
BS1188	-4.176	BH1627	0	lin1788	0	SAV0359	2	-	-	L0102	0	SP1525	0	cystathionine gamma-synthase (*yjcI*)
BS1313	-4.086	BH1505	1	-	-	-	-	-	-	L0117	0	SP0931	0	gamma-glutamyl kinase (*proB*)

BS3269	-3.989	BH3481	0	lin2514	1	SAV0837	1	CAC0984	1	L121289	0	SP0151	0	ABC transporter (ATP-binding protein) (*yusC*)
BS3900	-3.884	BH3232	0	-	-	-	-	CAC2807	0	-	-	-	-	endo-beta-1,3-1,4 glucanase (*bglS*)
BS0038	-3.883	BH0053	0	lin0216	0	SAV0490	0	CAC2991	1	L0353	0	SP0788	0	methionyl-tRNA synthetase (*metS*)
BS2858	-3.784	BH3111	1	lin1184	0	SAV1138	1	CAC2357	1	L0354	2	SP0579	1	phenylalanyl-tRNA synthethase (alpha subunit) (*pheS*)
BS0927	-3.776	BH1092	1	lin1574	1	SAV1300	1	CAC1319	1	L0015	0	SP2184	0	glycerol uptake facilitator (*glpF*)
BS1548	-3.473	BH2541	1	lin1954	1	SAV1198	1	CAC2113	1	L0227	1	SP1278	1	uracil phosphoribosyltransferase (*pyrR*)
BS3888	-3.412	-	-	lin0838	2	-	-	-	-	L124252	0	-	-	hypothetical protein (*yxjH*)
BS3889	-3.412	-	-	-	-	-	-	-	-	-	-	-	-	hypothetical protein (*yxjG*)
BS1549	-3.265	BH2540	1	lin1953	0	SAV1199	2	CAC2112	0	L46118	0	SP1286	0	uracil permease (*pyrP*)
BS1550	-3.120	BH2539	1	lin1952	0	SAV1200	0	CAC2654	1	L45002	0	SP1277	0	aspartate carbamoyltransferase (*pyrB*)
BS1166	-3.059	BH2679	0	lin0340	1	SAV2094	1	-	-	L0228	0	SP0722	0	transcriptional regulator (*tenA*)
BS1360	-2.966	-	-	-	-	-	-	-	-	-	-	-	-	ribulose-bisphosphate carboxylase (*ykrW*)
BS1855	-2.649	-	-	-	-	-	-	-	-	-	-	-	-	phosphoglycerate dehydrogenase (*yoaD*)
BS2376	-2.057	BH1503	0	lin0414	0	SAV1503	0	CAC3252	0	L135991	0	SP0933	0	pyrroline-5-carboxylate reductase (*yqjO*)

Table 4a. List of known attenuators in *B. subtilis* compared with predictions in six other genomes of gram-positive bacteria.

a "-" signifies that the absence of an ortholog.

b "-" signifies that no prediction is made because of absence of ortholog

c Gen back annotation of *B. subtilis* was used for gene definitions and gene names

Table 4b.

B. subtilis		*B. halodurans*		*L. innocua*		*S. aureus*		*C. acetobutylicum*		*L. lactis*		*S. pneumoniae*		
ID[a]	Z-score[b]	ID[a]	No. of predictions[c]	ID[a]	No. of predictions[c]	ID[a]	No. of predictions[c]	ID[a]	No. of predictions[c]	ID[a]	No. of predictions[c]	ID[a]	No. of predictions[c]	Gene[d]

BS1310	-6.836	BH0686	1	lin0846	0	-	-	-	-	-	-	-	-	chaperonin (*ykkC*)
BS3710	-5.196	BH3792	1	lin2704	1	SAV2127	0	CAC2892	0	L88252	1	SP0494	2	CTP synthetase (*ctrA*)
BS2484	-5.148	BH1417	0	lin1373	1	SAV1550	1	CAC1090	0	L172782	0	SP2095	0	5-formyltetrahydrofolate cyclo-ligase (*yqgN*)
BS0183	-4.938	-	-	-	-	SAV0452	1	-	-	-	-	-	-	NADH dehydrogenase (subunit 5) (*ndhF*)
BS3877	-4.696	-	-	lin1219	1	SAV0646	0	-	-	-	-	-	-	hypothetical protein (*yxkD*)
BS0535	-4.505	BH2084	1	-	-	-	-	-	-	-	-	-	-	antibiotic resistance protein (*ydfB*)
BS2452	-4.143	BH2816	0	lin1385	0	SAV1537	1	-	-	-	-	-	-	aminomethyltransferase (*yqhI*)
BS0143	-3.709	BH0162	1	lin2755	0	SAV2224	0	CAC3104	0	L0136	0	SP0236	0	RNA polymerase (alpha subunit) (*rpoA*)
BS0104	-3.612	BH0121	1	lin0282	1	SAV0539	0	CAC3146	0	L0407	0	SP1355	0	ribosomal protein L10 (*rplJ*)
BS0002	-3.480	BH0002	0	lin0002	0	SAV0002	0	CAC0002	1	L0275	0	SP0002	0	DNA polymerase III (beta subunit) (*dnaN*)
BS2881	-3.464	BH3140	1	lin1897	1	SAV1680	1	CAC2361	1	L95240	1	SP0959	0	initiation factor IF-3 (*infC*)
BS3093	-3.325	BH0830	0	lin1468	1	-	-	CAC2928	1	L116212	0	-	-	hypothetical protein (*yuaJ*)
BS2841	-3.262	BH3096	1	lin1198	0	SAV1393	1	-	-	-	-	-	-	aspartokinase II (*lysC*)
BS2903	-3.222	BH3153	0	lin1600	1	SAV1690	1	CAC1098	2	L0270	0	SP0032	0	DNA polymerase I (*polA*)
BS1920	-3.201	-	-	lin2900	0	SAV0721	0	CAC2687	1	L0268	0	-	-	ATP-dependent DNA helicase (*yocI*)
BS3731	-3.192	-	-	-	-	-	-	CAC2795	0	L19128	1	-	-	hypothetical protein (*ywiA*)
BS2278	-3.156	BH1643	1	-	-	-	-	-	-	-	-	-	-	hypothetical protein (*yphE*)
BS0878	-3.120	BH1933	1	-	-	-	-	CAC3014	1	-	-	-	-	thiamin biosynthesis protein (*thiA*)
BS2539	-3.075	BH1348	0	lin1509	0	SAV1579	1	CAC1283	1	L0272	1	SP0519	0	heat-shock protein (*dnaJ*)
BS1659	-2.908	BH2417	0	lin1358	1	SAV1265	0	CAC1798	1	L173151	1	SP0552	0	hypothetical protein (*ylxS*)
BS0115	-2.894	BH0133	1	lin2782	0	SAV2251	1	CAC3134	0	L0387	1	SP0208	0	ribosomal protein S10 (*rpsJ*)
BS0637	-2.878	BH0608	1	lin0582	0	SAV2253	0	CAC2772	0	-	-	-	-	hypothetical protein (*yebB*)
BS2627	-2.867	-	-	lin0073	0	SAV1098	1	-	-	-	-	-	-	transcriptional regulator (*yqaE*)
BS2986	-2.643	BH3263	0	lin1657	0	SAV1749	1	-	-	L155396	0	SP0549	0	hypothetical protein (*ytmP*)
BS1659	-2.573	BH2417	0	lin1358	1	SAV1265	0	CAC1798	1	L173151	1	SP0552	0	hypothetical protein (*ylxS*)
BS0636	-2.553	BH0607	1	lin1081	0	SAV0391	0	CAC2700	0	L115968	1	SP1445	0	GMP synthetase (*guaA*)
BS0558	-2.516	-	-	lin1827	1	-	-	-	-	L53789	1	-	-	hypothetical protein (*ydgC*)
BS2512	-2.483	BH1379	0	lin1490	0	SAV1560	1	CAC1302	0	L84257	0	SP1610	0	hypothetical protein (*yqfN*)
BS0561	-2.386	-	-	-	-	-	-	CAC3007	2	-	-	-	-	ATP-binding transport protein (*expZ*)
BS3473	-2.308	BH3570	0	lin1215	1	SAV0467	0	CAC0084	0	L0288	0	SP0740	0	mutator protein (*yvcI*)
BS1319	-2.296	BH0438	1	lin1789	1	SAV0356	0	-	-	L0100	0	SP0585	0	cobalamin-independent methionine synthase (*metC*)
BS3265	-2.287	BH3471	0	lin2510	0	SAV0842	0	CAC3288	1	L34806	0	SP0867	1	ABC transporter (ATP-binding protein) (*yurY*)
BS0668	-2.210	BH0665	0	lin1868	0	SAV1901	0	CAC2671	1	L0475	0	SP0438	0	glutamyl-tRNA(Gln) amidotransferase (subunit C) (*yerL*)
BS2729	-2.163	BH1272	0	lin1533	0	SAV1614	1	CAC1686	0	L165684	0	SP0980	0	*O*-methyltransferase (*yrrM*)
BS1386	-2.151	BH0744	0	lin0644	1	-	-	-	-	L63697	0	-	-	heavy metal-transporting ATPase (*ykvW*)

BS0019	-	BH0034	1	lin2852	0	SAV0478	0	CAC0125	1	L0279	0	SP0865	0	DNA polymerase III (gamma and tau subunits) (*dnaX*)
BS0106	-	BH0124	1	lin0284	0	SAV0541	0	CAC0877	0	L65498	0	SP0841	1	hypothetical protein (*ybxB*)
BS0107	-	BH0126	0	lin0285	1	SAV0542	1	CAC3143	1	L0137	1	SP1961	1	DNA-directed RNA polymerase (beta subunit) (*rpoB*)
BS0112	-	BH0131	0	lin2803	0	SAV0547	0	CAC3138	1	L0368	1	SP0273	0	elongation factor G (*fus*)
BS0462	-	BH0518	0	lin0884	0	SAV2071	0	CAC0489	1	L61355	1	SP1699	0	acyl carrier protein synthase (*ydcB*)
BS0516	-	BH0725	0	-	-	SAV2534	1	CAC0875	1	L124727	0	SP1447	0	hypothetical protein (*ydeD*)
BS0663	-	BH0649	0	lin1870	0	SAV1904	0	CAC2673	0	L0304	1	SP1117	1	DNA ligase (*yerG*)
BS0692	-	-	-	lin0655	0	SAV1419	0	CAC2751	1	L36177	1	SP1464	0	hypothetical protein (*yesJ*)
BS0824	-	BH3305	0	lin0649	1	SAV2522	0	-	-	-	-	SP0073	1	hypothetical protein (*yfiE*)
BS0892	-	BH1023	0	-	-	SAV1855	0	CAC0700	0	L161988	1	SP0486	1	rRNA methylase (*cspR*)
BS0960	-	BH2987	1	-	-	SAV1783	0	CAC1586	1	-	-	SP1295	0	hypothetical protein (*yhdV*)
BS0988	-	BH0202	0	lin1781	0	SAV1045	0	CAC2712	1	L0171	1	SP0415	1	similar to 3-hydroxybutyryl-CoA dehydratase (*yhaR*)
BS1062	-	-	-	lin2369	0	SAV0966	0	CAC2263	1	L0252	1	SP1151	0	ATP-dependent deoxyribonuclease (subunit B) (*addB*)
BS1139	-	BH3636	0	lin0182	1	SAV0991	0	CAC3179	1	L88446	0	-	-	oligopeptide ABC transporter (oligopeptide-binding protein) (*appA*)
BS1269	-	BH0012	1	lin2568	0	SAV1955	1	CAC1883	0	L60836	0	-	-	prophage (*xkd0*)
BS1345	-	BH2553	1	lin0990	1	SAV1022	1	CAC1415	0	-	-	-	-	similar to toxic anion resistance protein (*ykoY*)
BS1390	-	BH0844	1	-	-	SAV0189	1	CAC0570	0	-	-	SP1684	0	PTS glucose-specific enzyme II ABC component (*ptsG*)
BS1478	-	BH2632	0	lin1055	1	SAV1109	0	CAC1684	0	L0370	1	SP0681	0	similar to GTP-binding elongation factor (*ylaG*)
BS1512	-	BH2579	0	lin2152	1	SAV2443	2	CAC2937	0	L157055	1	-	-	similar to ketopantoate reductase (*ylbQ*)
BS1514	-	BH2576	1	lin2148	1	SAV1178	0	CAC2133	0	-	-	-	-	hypothetical protein (*yllB*)
BS1553	-	BH2536	0	lin1949	0	SAV1203	0	CAC2644	0	L198033	2	SP1275	1	carbamoyl-phosphate synthase (catalytic subunit) (*pyrAB*)
BS1566	-	BH2515	1	lin0832	0	-	-	CAC2137	0	L2866	1	SP1551	1	cation-transporting ATPase (*yloB*)
BS1570	-	BH2510	0	lin1939	1	SAV1211	1	CAC1720	0	L166912	0	SP1231	0	pantothenate metabolism flavoprotein (*yloI*)
BS1587	-	BH2495	0	lin1925	0	SAV1227	0	CAC1736	1	L0262	1	SP1697	0	ATP-dependent DNA helicase *recG* (*ylpB*)
BS1614	-	BH1529	0	lin1316	0	SAV1252	0	CAC2066	0	L34517	1	SP0890	1	integrase/recombinase (*codV*)
BS1650	-	BH2426	0	lin1766	0	SAV1257	1	CAC1788	0	L0376	1	SP2214	0	elongation factor Ts (*tsf*)
BS1653	-	BH2423	1	lin1352	1	SAV1260	1	CAC1791	0	L183602	0	SP0261	1	undecaprenyl diphosphate synthetase (*yluA*)
BS1668	-	BH2408	0	lin1367	1	SAV1273	0	CAC1807	1	L0392	0	SP1626	0	ribosomal protein S15 (*rpsO*)
BS1669	-	BH2407	0	lin1368	1	SAV1274	1	CAC1808	0	L0325	0	SP0588	1	polyribonucleotide phosphorylase (*pnpA*)

BS1677	-	BH1742	1	lin1474	0	SAV1395	0	CAC2378	0	L0093	1	SP1014	0	dihydrodipicolinate synthase (*dapA*)
BS2023	-	BH3508	0	-	-	-	-	CAC1501	1	-	-	SP1336	1	DNA-methyltransferase (Bsu) (*mtbP*)
BS2185	-	BH3062	0	lin2090	1	SAV2053	1	CAC3170	0	L0077	0	SP2126	1	dihydroxy-acid dehydratase (*ilvD*)
BS2258	-	BH1665	0	lin2039	0	SAV0724	1	CAC3031	1	L0065	0	-	-	histidinol-phosphate aminotransferase (*hisC*)
BS2264	-	BH1659	1	lin1674	1	SAV1367	1	CAC3163	1	L0054	1	SP1817	1	anthranilate synthase (*trpE*)
BS2301	-	-	-	lin2059	1	SAV1485	0	CAC2841	1	L106755	0	SP0488	0	hypothetical protein (*ypaA*)
BS2324	-	BH1554	1	-	-	SAV1771	1	CAC0590	0	L0163	2	SP0178	2	riboflavin-specific deaminase (*ribG*)
BS2334	-	BH1544	1	lin2066	0	SAV1400	0	CAC0608	1	L0121	0	SP1978	0	diaminopimelate decarboxylase (*lysA*)
BS2383	-	BH1472	0	lin2082	0	SAV1895	1	CAC0285	0	L0305	0	SP0458	1	DNA-damage inducible protein (*yqjH*)
BS2661	-	-	-	-	-	SAV1407	0	CAC1610	1	-	-	SP0626	1	branched-chain amino acid transporter (*brnQ*)
BS2741	-	BH1262	1	lin1545	0	SAV1620	1	CAC1067	0	-	-	-	-	hypothetical protein (*yrrB*)
BS2747	-	BH1255	1	-	-	SAV1628	1	CAC0908	0	-	-	SP0695	0	hypothetical protein (*yrvM*)
BS2932	-	BH0170	0	lin2443	0	SAV2414	0	CAC3325	1	L162009	1	SP0148	0	similar to amino acid ABC transporter (binding protein) (*ytmJ*)
BS2942	-	BH3193	0	lin1617	0	SAV1712	1	-	-	L74738	1	SP2045	1	hypothetical protein (*ytxK*)
BS3015	-	BH0783	1	-	-	SAV2427	1	CAC1361	0	-	-	-	-	dethiobiotin synthetase (*bioD*)
BS3049	-	BH3300	1	lin1773	1	SAV1790	1	CAC2856	1	L153408	1	SP0762	0	*S*-adenosylmethionine synthetase (*metK*)
BS3345	-	BH0557	0	lin1967	0	SAV2557	1	CAC3655	0	L45966	1	SP0729	0	heavy metal-transporting ATPase (*yvgX*)
BS3388	-	BH3559	0	lin2552	1	SAV0773	1	CAC0710	1	L0010	0	SP0499	0	3-phosphoglycerate kinase (*pgk*)
BS3496	-	BH0421	0	lin0955	1	SAV0701	1	CAC0188	0	L173068	0	SP2056	0	*N*-acetylglucosamine-6-phosphate deacetylase (*nagA*)
BS3705	-	BH3784	0	lin2697	0	SAV2124	0	CAC3539	1	L113067	1	SP1081	0	UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase (*murZ*)
BS3728	-	BH0834	1	lin2706	0	SAV0607	0	CAC1041	1	L0344	0	SP2078	0	arginyl-tRNA synthetase (*argS*)
BS3729	-	BH3809	1	lin2707	0	SAV0606	1	CAC2894	0	-	-	-	-	hypothetical protein (*ywiB*)
BS3875	-	BH1141	0	lin2875	1	-	-	CAC3236	0	L26721	1	SP0593	0	hypothetical protein (*yxkF*)
BS3901	-	BH0297	0	lin2530	1	SAV1357	0	CAC0422	1	L0154	1	SP0576	0	transcriptional antiterminator (BglG family) (*licT*)
BS3916	-	-	-	lin0454	1	-	-	CAC1057	1	-	-	-	-	cell wall-associated protein precursor (*wapA*)
BS3918	-	BH3184	1	lin1615	0	SAV1710	0	-	-	L84477	1	SP1996	0	hypothetical protein (*yxiE*)
BS3919	-	-	-	lin0344	1	-	-	CAC0743	1	-	-	SP0578	0	beta-glucosidase (*bglH*)
BS4099	-	BH4065	0	lin2987	0	SAV2713	1	CAC3738	1	L131443	0	SP2042	0	ribonuclease P (protein component) (*rnpA*)
-	-	BH0440	2	-	-	SAV0142	1	-	-	L97415	0	-	-	phosphonate ABC transporter ATP-binding protein
-	-	BH0503	0	-	-	-	-	CAC2487	1	L3279	0	SP0204	1	hypothetical protein
-	-	BH0595	1	lin0026	1	-	-	CAC1407	1	L90678	0	-	-	PTS beta-glucoside-specific enzyme

II ABC component														
-	-	BH2025	1	lin2903	1	-	-	-	-	-	-	-	-	ABC transporter (ATP-binding protein)
-	-	BH2200	1	-	-	-	-	-	-	L0358	1	SP2229	0	tryptophanyl-tRNA synthetase
-	-	BH2603	0	lin0604	1	-	-	CAC2783	1	L75975	0	-	-	*O*-acetylhomoserine sulfhydrylase
-	-	-	-	-	-	-	-	CAC1072	1	L180742	1	-	-	hypothetical protein
-	-	-	-	-	-	-	-	-	-	L49741	1	SP1647	1	endopeptidase
-	-	-	-	-	-	-	-	-	-	L90622	1	SP0861	1	hypothetical protein
-	-	-	-	lin1136	1	-	-	CAC2721	1	-	-	-	-	similar to two-component response regulator
-	-	-	-	lin1443	1	SAV0226	0	CAC0980	1	L57408	0	SP0459	0	pyruvate-formate lyase
-	-	-	-	lin1851	1	-	-	CAC3619	0	-	-	SP1502	1	similar to amino acid (glutamine) ABC transporter, permease protein
-	-	-	-	lin2124	1	SAV1402	1	CAC2990	0	-	-	-	-	major cold-shock protein

Table 4b. List of all orthologous genes in the six gram-positive bacteria genomes in which two or more genomes share predicted attenuators.

a "-" signifies that the absence of an ortholog.

b "-" signifies that no prediction is made because of absence of ortholog

c "-" signifies that no prediction is made because of absence of ortholog

d GenBank annotation of *B. subtilis* was used for gene definitions and gene names. If ortholog is missing in *B. subtilis*, GenBank annotation of on of the other six genomes was used for gene definitions.

Table 5a.

E. coli		*H. influenzae*		*V. cholerae*		*P. aeruginosa*		*X. fastidiosa*		
ID	Z-score	ID[a]	No. of predictions[b]	ID[a]	No. of predictions[b]	ID[a]	No. of predictions[b]	ID[a]	No. of predictions[b]	Gene[c]
b2019	-7.560	HI0468	1	VC1132	1	PA4449	0	XF2220	0	ATP phosphoribosyltransferase (*hisG*)
b3767	-4.529	-	-	-	-	-	-	-	-	acetolactate synthase II, large subunit (*ilvG*)
b2599	-3.945	HI1145	0	VC0705	0	PA3166	0	XF2325	0	chorismate mutase-P and prephenate dehydratase (*pheA*)
b3671	-3.923	-	-	VC0031	0	-	-	XF1821	0	acetolactate synthase I, large subunit (*ilvB*)
b3722	-3.823	-	-	-	-	-	-	-	-	PTS beta-glucosides, enzyme II (*bglF*)
b0002	-3.788	HI0089	1	VC2364	1	PA0904	0	XF2225	0	aspartokinase I homoserine dehydrogenase I (*thrA*)
b3752	-3.573	HI0505	0	VCA0131[d]	-	PA1950	0	XF0366	0	ribokinase (*rbsK*)
b0074	-3.559	HI0986	2	VC2490	1	PA1217	0	XF1818	0	2-isopropylmalate synthase (*leuA*)
b0170	-3.506	HI0914	1	VC2259	1	PA3655	0	XF2579	0	protein chain elongation factor EF-Ts (*tsf*)

b3813	-3.503	HI1188	0	VC0190	0	PA5443	0	XF0050	0	DNA-dependent ATPase I and helicase II (*uvrD*)
b4245	-3.387	-	-	VC2510	0	PA0402	0	XF2226	0	aspartate carbamoyltransferase, catalytic subunit (*pyrB*)
b1264	-2.717	HI1387	0	VC1174	1	PA1001	0	-	-	anthranilate synthase component I (*trpE*)
b3642	-2.675	HI0272	0	VC0211	1	PA5331	0	XF0153	0	orotate phosphoribosyltransferase (*pyrE*)
b3723	-2.565	-	-	-	-	-	-	-	-	positive regulation of *bgl* operon (*bglG*)

Table 5a. List of known attenuators in *E. coli* compared with predictions in four other genomes of proteobacteria (gamma subdivision).

a "-" signifies that the absence of an ortholog.

b "-" signifies that no prediction is made because of absence of ortholog

c GenBank annotation of *E. coli* was used for gene definitions and gene names

Table 5b.

E. coli		*H. influenzae*		*V. cholerae*			*P. aeruginosa*			*X. fastidiosa*			
ID	Z-score[a]		ID[b]	No. of predictions[c]	ID[b]		No. of predictions[c]	ID[b]		No. of predictions[c]	ID[b]	No. of predictions[c]	Gene[d]
b3828	-7.143	HI1739	1	VC1706	0	PA3587	0	-	-	regulator for *metE* and *metH* (*metR*)			
b2425	-6.168	-	-	VC0538	1	-	-	-	-	thiosulfate binding protein (*cysP*)			
b3066	-5.002	HI0532	1	VC0518	-	PA0577	0	XF0430	0	DNA primase (*dnaG*)			
b0902	-4.443	HI0179	1	VC1869	0	PA1919	0	-	-	pyruvate formate lyase activating enzyme 1 (*pflA*)			
b3871	-4.352	HI0864	1	VC2744	0	PA5117	0	XF1213	0	putative GTP-binding factor (*yihK*)			
b3181	-3.889	HI1331	1	VC0634	0	PA4755	0	XF1108	0	transcription elongation factor (*greA*)			
b3983	-3.861	HI0517	1	VC0324	0	PA4274	0	XF2637	0	ribosomal protein L11 (*rplK*)			

b0610	-3.506	-	-	-	-	PA5274	1	-	-	regulator of nucleoside diphosphate kinase (*rnk*)
b3298	-3.362	HI0799	0	VC2574	1	PA4241	0	XF1173	0	ribosomal protein S13 (*rpsM*)
b0680	-3.260	HI1354	1	VC0997	0	PA1794	0	XF1338	0	glutamine tRNA synthetase (*glnS*)
b2313	-2.990	HI1206	1	VC1003	0	PA3109	0	XF1948	0	membrane protein required for colicin V production (*cvpA*)
b0441	-2.961	HI1004	0	VC1918	1	PA1805	0	XF1191	0	putative protease maturation protein (*ybaU*)
b1253	-2.808	HI0827	0	VC1701	0	PA5371	1	-	-	hypothetical protein (*yciA*)
b2924	-2.487	-	-	VC0480	0	PA4394	1	XF1258	0	putative transport protein (*yggB*)
b2140	-2.289	HI0270	0	VC1105	0	PA3129	1	-	-	putative regulator protein (*yohI*)
b1778	-2.121	HI1455	1	VC1998	0	PA2827	0	XF0849	0	hypothetical protein (*yeaA*)
b2185	-	HI1630	1	VC1640	1	PA4671	1	XF2643	0	ribosomal protein L25 (*rplY*)
b2944	-	HI1173	1	VC0471	1	PA1189	0	-	-	hypothetical protein (*sprT*)
b3170	-	HI1282	1	VC0641	0	PA4746	0	XF0233	1	hypothetical protein (*yhbC*)
b3936	-	HI0758	1	VC2679	1	PA5049	0	XF1534	0	ribosomal protein L31 (*rpmE*)
b3987[e]	-	HI0515	1	VC0328	1	PA4270	1	XF2633	0	DNA-directed RNA polymerase (beta subunit) (*rpoB*)
b4006	-	HI0887	1	VC0276	1	PA4854	0	XF1975	0	phosphoribosylaminoimidazolecarboxamide formyltransferase (*purH*)

Table 5b. List of all orthologous genes in the five proteobacteria (gamma subdivision) genomes in which two or more genomes share predicted attenuators.

a "-" signifies that the absence of an ortholog.

b "-" signifies that no prediction is made because of absence of ortholog

c "-" signifies that no prediction is made because of absence of ortholog

d GenBank annotation of *B. subtilis* was used for gene definitions and gene names. If ortholog is missing in *B. subtilis*, GenBank annotation of on of the other six genomes was used for gene definitions.

e An attenuator is known for this gene but our method did not predict predict an attenuator.

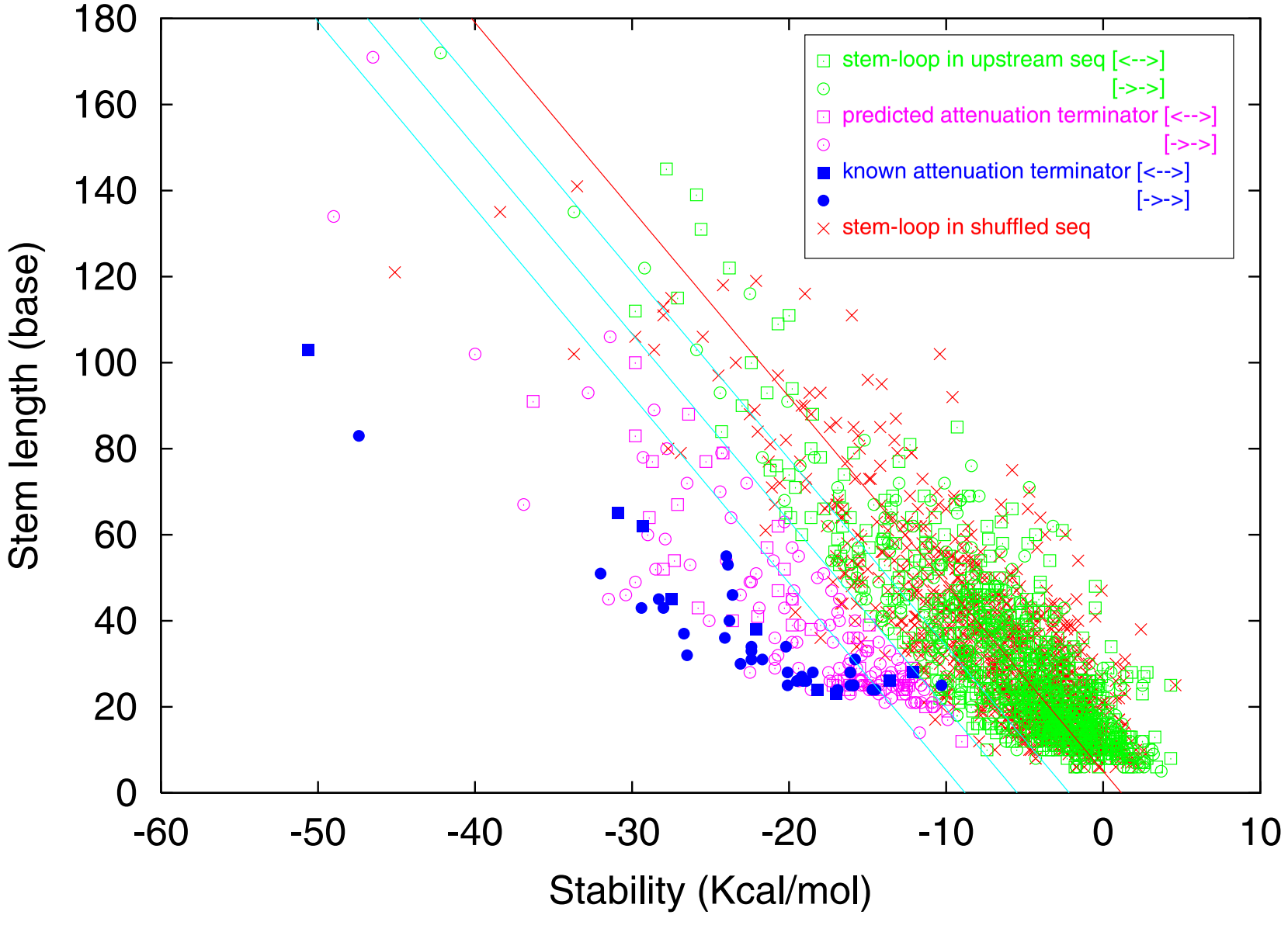

Stem length (base)
Stability (Kcal/mol)
180
160
140
120
100
80
60
40
20
0
-60
-50
-40
-30
-20
-10
0
10
stem-loop in upstream seq [<-->]
[->->]
predicted attenuation terminator [<-->]
[->->]
known attenuation terminator [<-->]
[->->]
stem-loop in shuffled seq

Figure 1

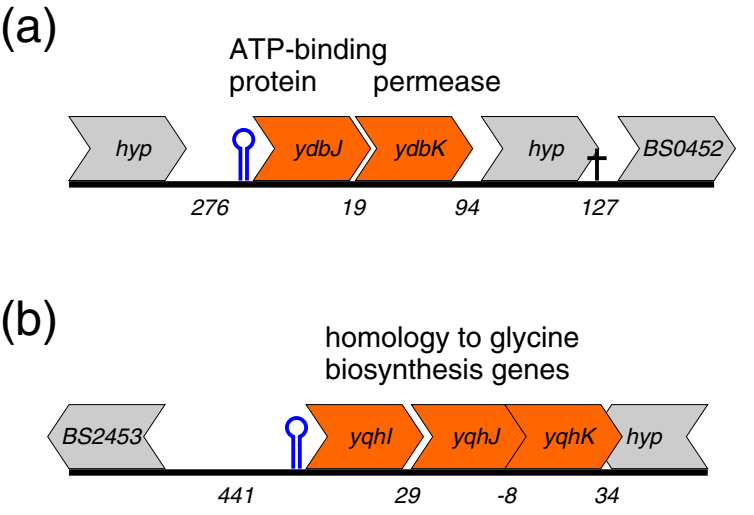

(a)
ATP-binding
protein
permease
hyp
ydbJ
ydbK
hyp
BS0452
276
19
94
127
(b)
homology to glycine
biosynthesis genes
BS2453
yqhI
yqhJ
yqhK
hyp
441
29
-8
34

Figure 2

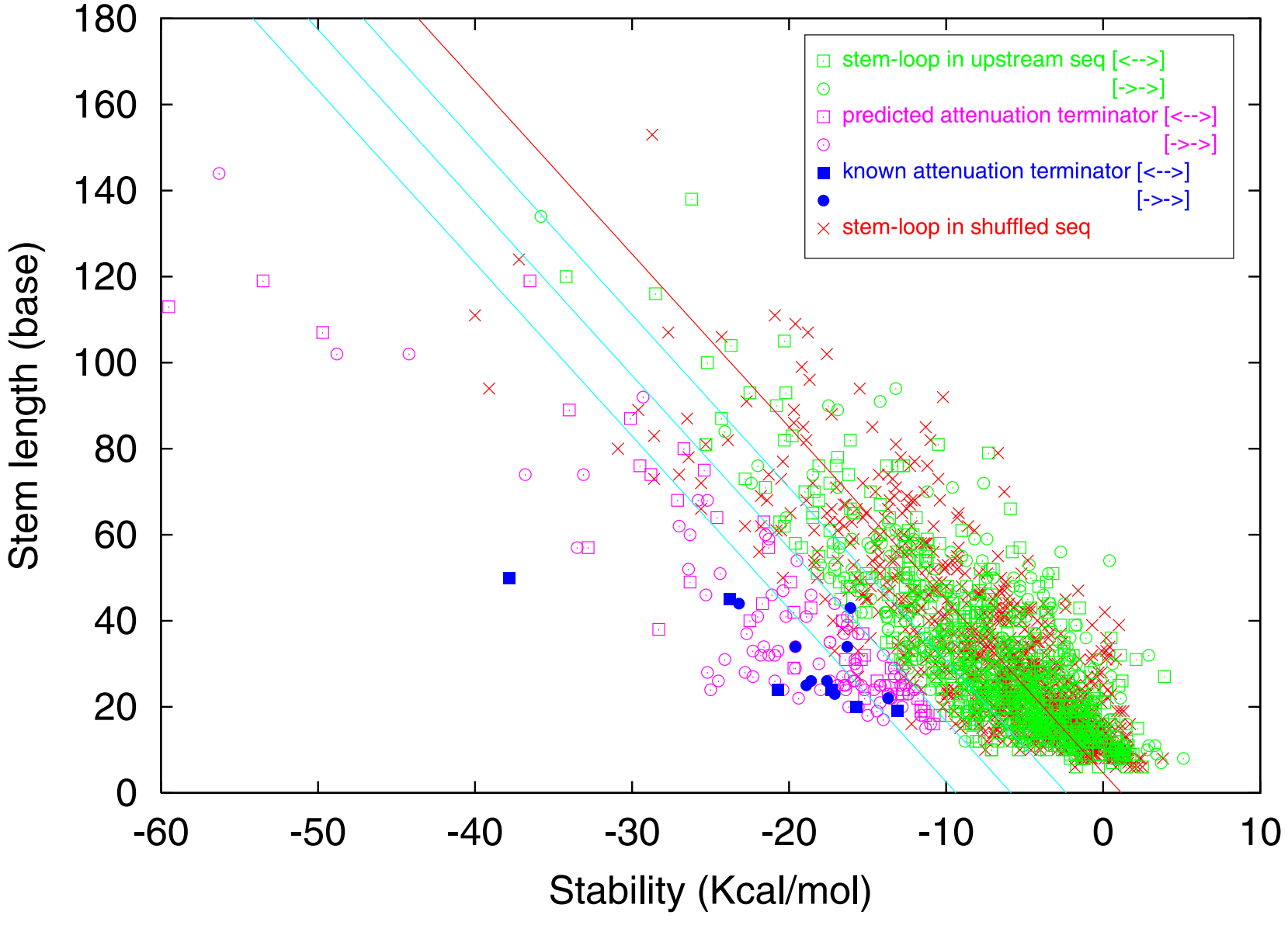

Stem length (base)
Stability (Kcal/mol)
180
160
140
120
100
80
60
40
20
0
-60
-50
-40
-30
-20
-10
0
10
stem-loop in upstream seq [<-->]
[->->]
predicted attenuation terminator [<-->]
[->->]
known attenuation terminator [<-->]
[->->]
stem-loop in shuffled seq

Figure 3

Figure 4: Putative attenuators vs intergenic

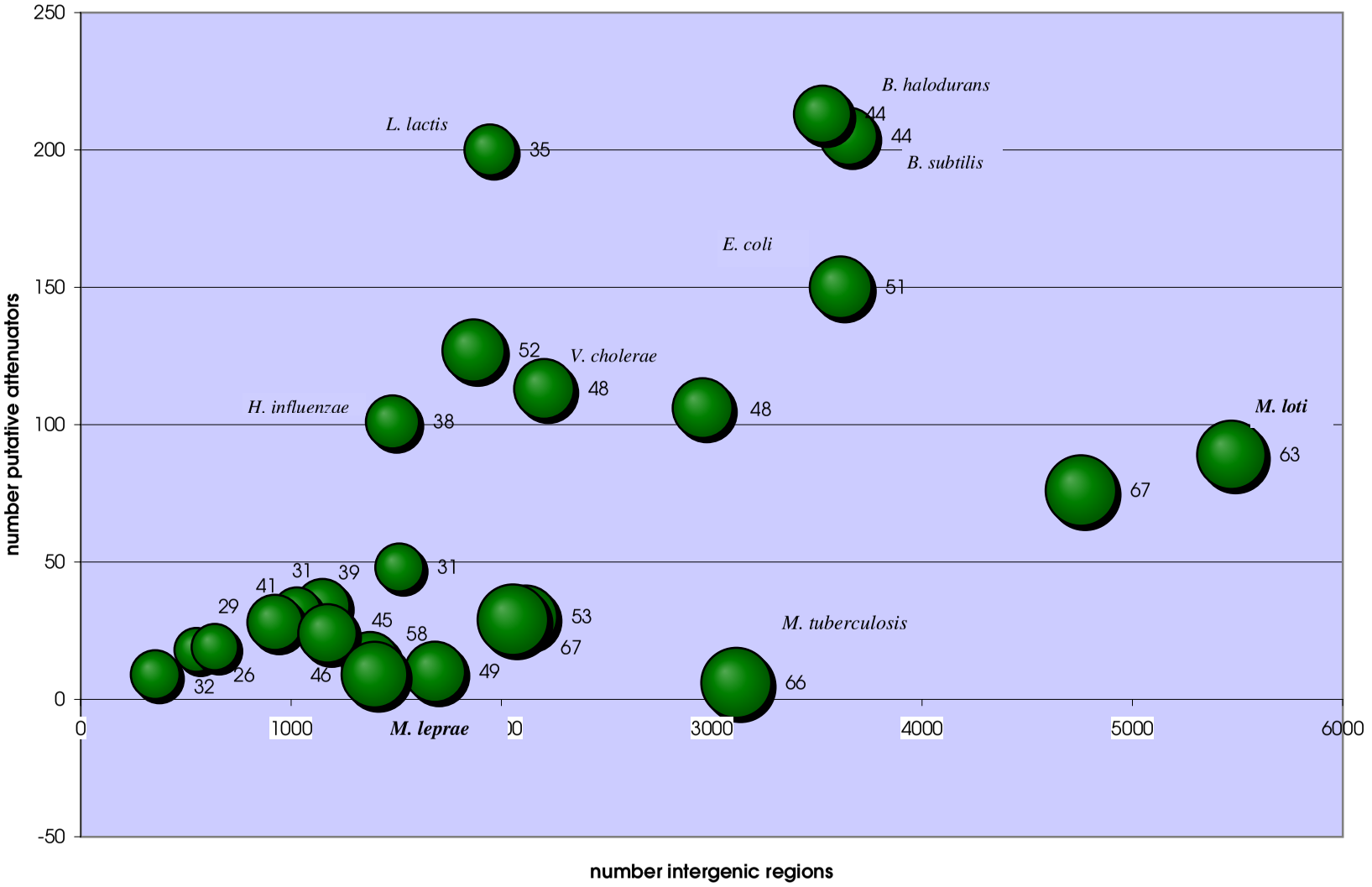


Figure 5a: Genome size and random folds

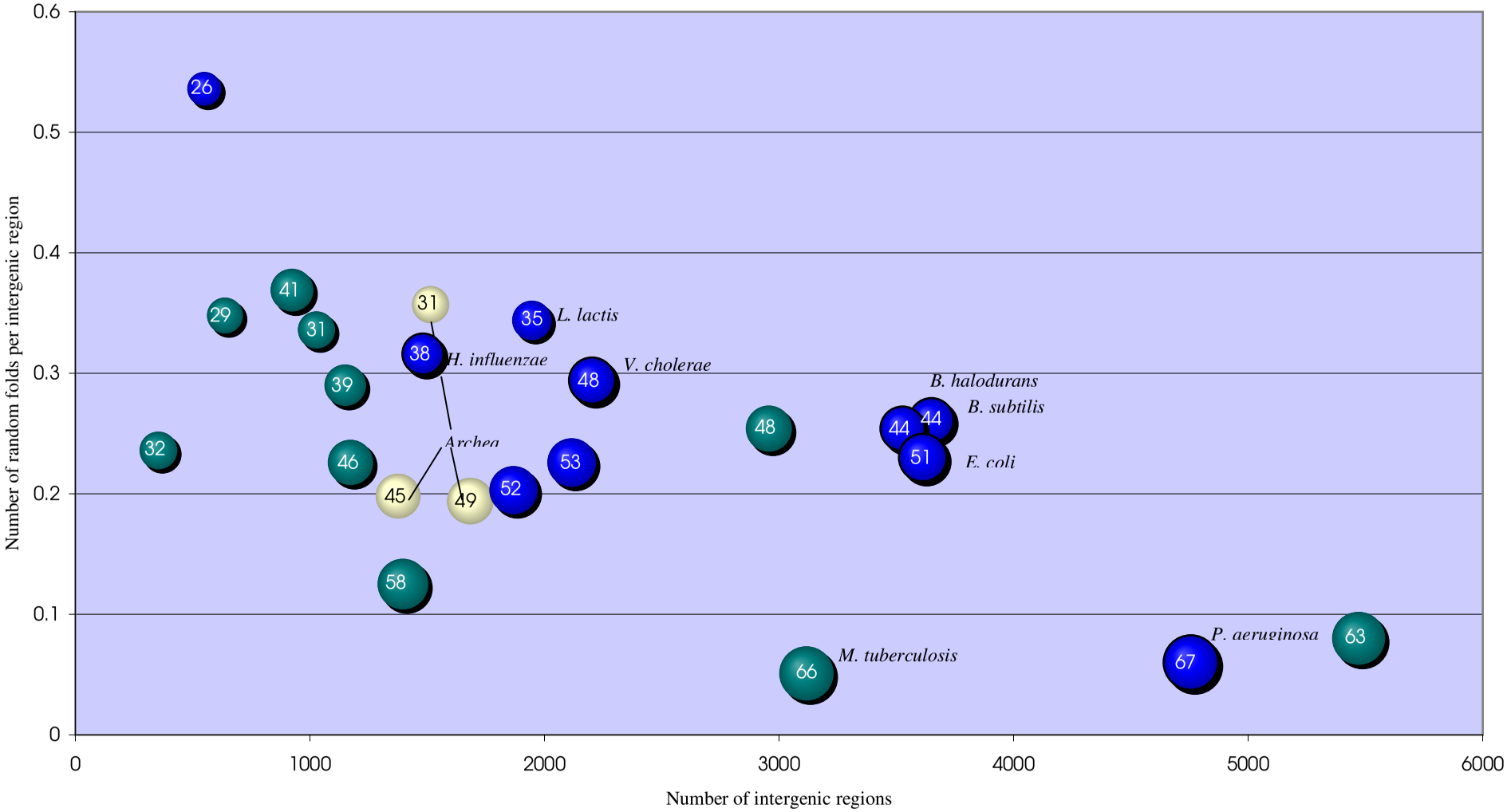
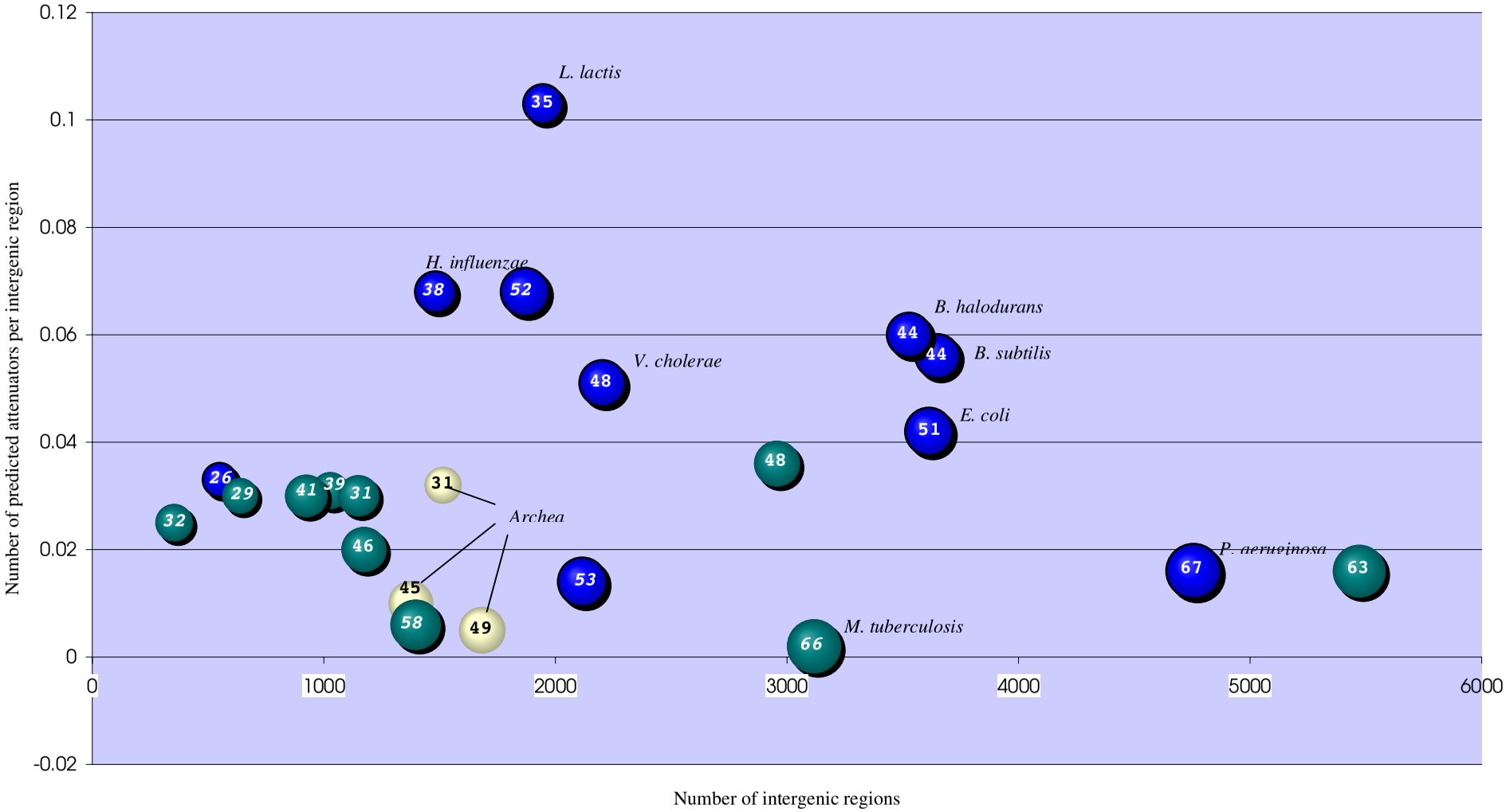


Figure 5b: Genome Size and Attenuators



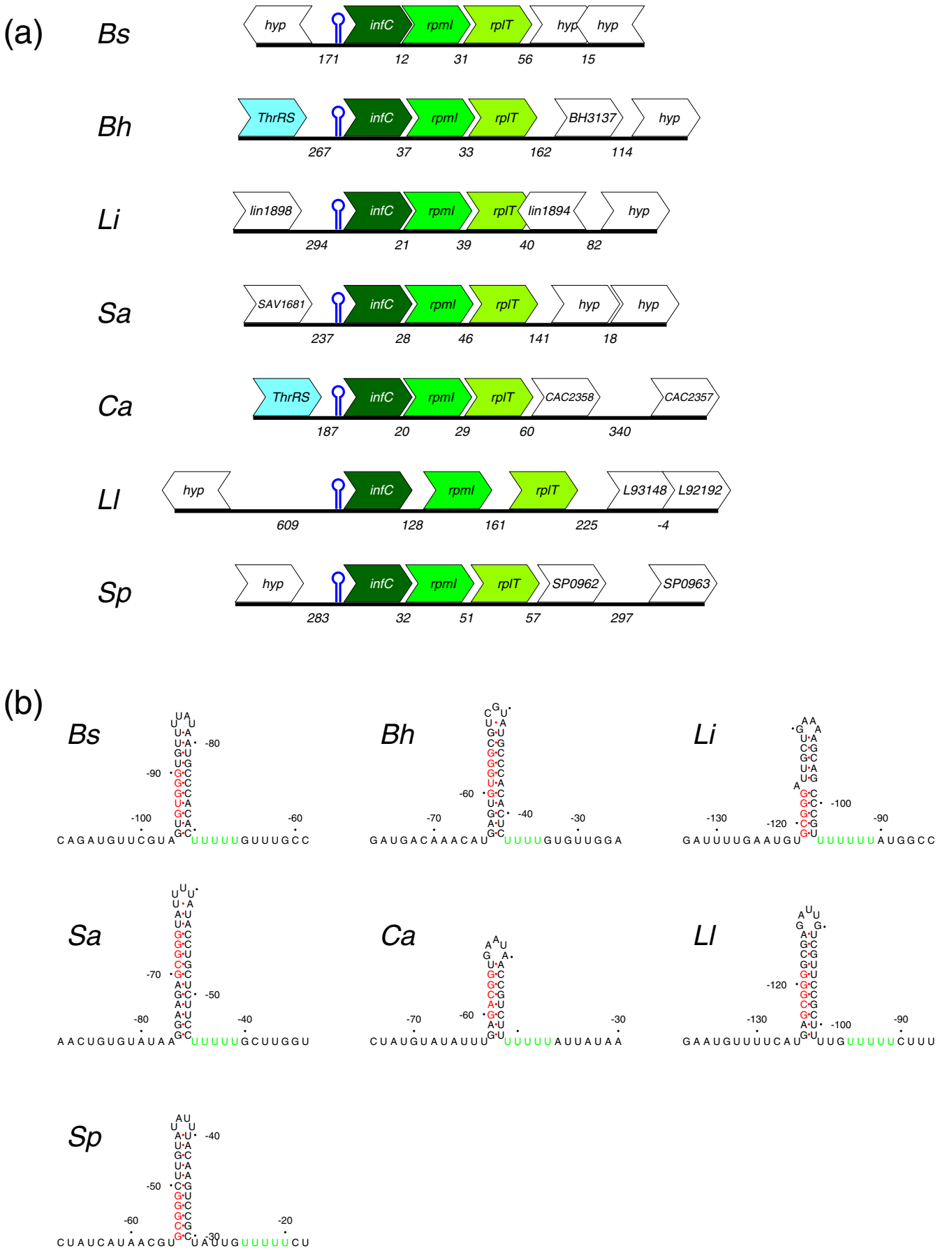

(a)
Bs
hyp
infC
rpmI
rplT
hyp
hyp
171
12
31
56
15
Bh
ThrRS
infC
rpmI
rplT
BH3137
hyp
267
37
33
162
114
Li
lin1898
infC
rpmI
rplT
lin1894
hyp
294
21
39
40
82
Sa
SAV1681
infC
rpmI
rplT
hyp
hyp
237
28
46
141
18
Ca
ThrRS
infC
rpmI
rplT
CAC2358
CAC2357
187
20
29
60
340
Ll
hyp
infC
rpmI
rplT
L93148
L92192
609
128
161
225
-4
Sp
hyp
infC
rpmI
rplT
SP0962
SP0963
283
32
51
57
297
(b)
Bs
CAGAUGUUCGUA
UUUUU
GUUUGCC
Bh
GAUGACAAACAU
UUUU
GUGUUGGA
Li
GAUUUUGAAUGU
UUUUUU
AUGGCC
Sa
AACUGUGUAUAA
UUUUU
GCUUGGU
Ca
CUAUGUAUAUUU
UUUUU
AUUAUAA
Ll
GAAUGUUUUCAU
UUG
UUUUUU
CUUU
Sp
CUAUCAUAACGU
UAUUG
UUUUU
CU

Figure 6

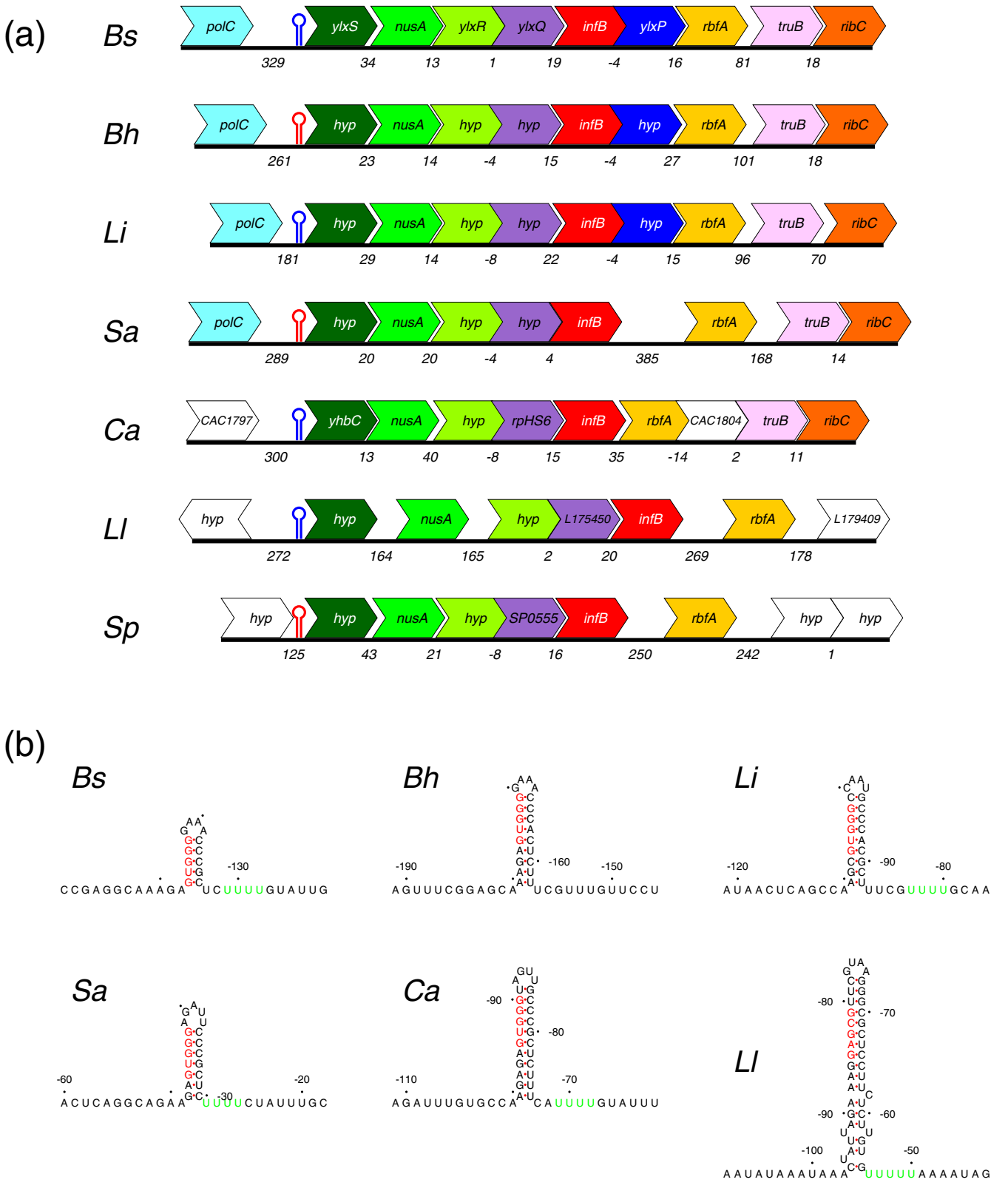

(a)
Bs
polC
ylxS
nusA
ylxR
ylxQ
infB
ylxP
rbfA
truB
ribC
329
34
13
1
19
-4
16
81
18
Bh
polC
hyp
nusA
hyp
hyp
infB
hyp
rbfA
truB
ribC
261
23
14
-4
15
-4
27
101
18
Li
polC
hyp
nusA
hyp
hyp
infB
hyp
rbfA
truB
ribC
181
29
14
-8
22
-4
15
96
70
Sa
polC
hyp
nusA
hyp
hyp
infB
rbfA
truB
ribC
289
20
20
-4
4
385
168
14
Ca
CAC1797
yhbC
nusA
hyp
rpHS6
infB
rbfA
CAC1804
truB
ribC
300
13
40
-8
15
35
-14
2
11
Ll
hyp
hyp
nusA
hyp
L175450
infB
rbfA
L179409
272
164
165
2
20
269
178
Sp
hyp
hyp
nusA
hyp
SP0555
infB
rbfA
hyp
hyp
125
43
21
-8
16
250
242
1
(b)
Bs
-130
CCGAGGCAAAGAUCUUUUGUAUUG
Bh
-190
-160
-150
AGUUUCGGAGCAUCGUUUGUUCCU
Li
-120
-90
-80
AUAACUCAGCCAUUCGUUUUGCAA
Sa
-60
-30
-20
ACUCAGGCAGAAUUUUCUAUUUGC
Ca
-90
-80
-110
-70
AGAUUUGUGCCAUCAUUUUGUAUUU
Ll
-90
-80
-70
-100
-50
AAUAUAAAUAAAUUUUUUAAAAUAG
Sp
-30
-20
-50
-10
AAUAUGAAAAGUUUGCUUAUGAGG

Figure 7

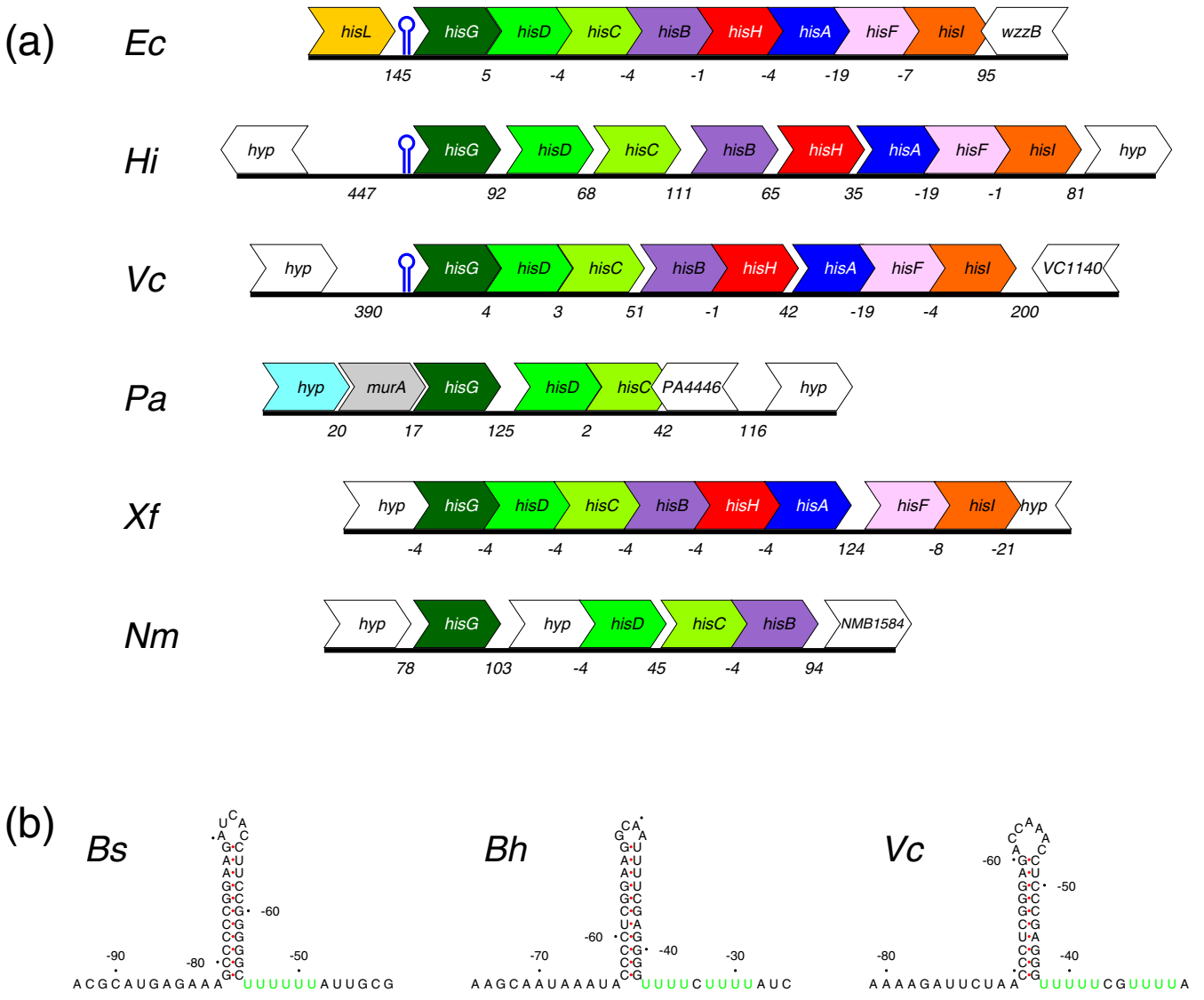

(a)
Ec
hisL
hisG
hisD
hisC
hisB
hisH
hisA
hisF
hisI
wzzB
145 5 -4 -4 -1 -4 -19 -7 95
Hi
hyp
hisG
hisD
hisC
hisB
hisH
hisA
hisF
hisI
hyp
447 92 68 111 65 35 -19 -1 81
Vc
hyp
hisG
hisD
hisC
hisB
hisH
hisA
hisF
hisI
VC1140
390 4 3 51 -1 42 -19 -4 200
Pa
hyp
murA
hisG
hisD
hisC
PA4446
hyp
20 17 125 2 42 116
Xf
hyp
hisG
hisD
hisC
hisB
hisH
hisA
hisF
hisI
hyp
-4 -4 -4 -4 -4 -4 124 -8 -21
Nm
hyp
hisG
hyp
hisD
hisC
hisB
NMB1584
78 103 -4 45 -4 94
(b)
Bs
-90
-80
-60
-50
ACGCAUGAGAAA UUUUUU AUUGCG
Bh
-70
-60
-40
-30
AAGCAAUAAAUA UUUU CUUUU AUC
Vc
-80
-60
-50
-40
AAAAGAUUCUAA UUUUU CGUUUUA

Figure 8

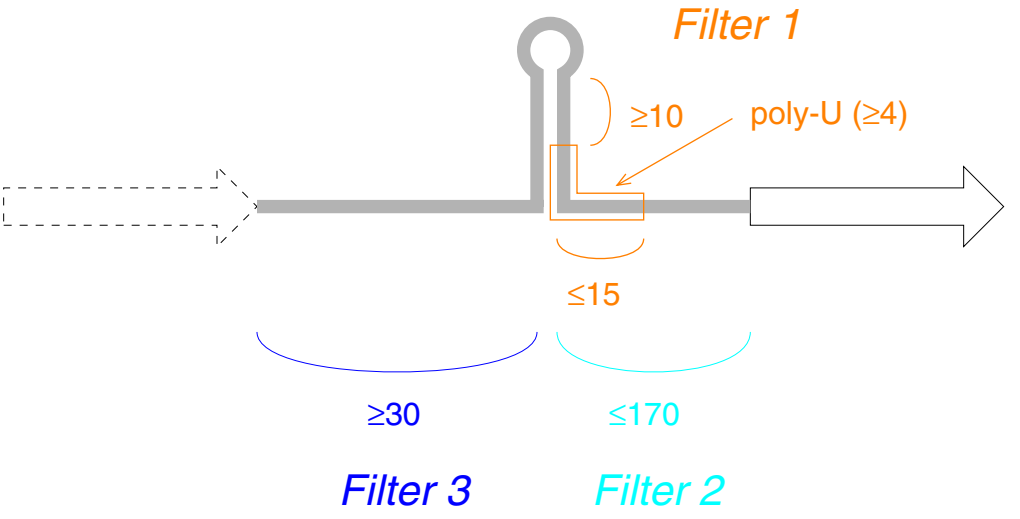


Figure 9