Tuberculosis (2008) **88**, 566–575

available at www.sciencedirect.com

ScienceDirect

journal homepage: http://intl.elsevierhealth.com/journals/tube

Genome-wide analysis of the intrinsic terminators of transcription across the genus *Mycobacterium*

A. Mitra [a], K. Angamuthu [a,c], V. Nagaraja [a,b,*]

[a] *Department of Microbiology and Cell Biology, Indian Institute of Science, C.V. Raman Avenue, Bangalore 560012, India*
[b] *Jawaharlal Nehru Center for Advanced Scientific Research, Bangalore 560064, India*

Received 20 January 2008; received in revised form 8 June 2008; accepted 10 June 2008

KEYWORDS
Transcription termination;
RNA polymerase;
Mycobacterium;
GeSTer

Summary
Termination of transcription in eubacteria is achieved by a region of the nascent transcript. In *Escherichia coli*, this intrinsic terminator consists of a hairpin followed by a U-stretch. Absence of the typical terminators in several genes of *Mycobacterium tuberculosis* led us to develop an accurate and efficient algorithm to identify putative terminators in all sequenced microbial genomes. In addition to the typical *Escherichia coli* type of terminators, several variant terminator structures were predicted by the algorithm and their existence was experimentally verified. We have now analysed 17 *Mycobacterium* genomes to obtain a comprehensive picture of the transcription terminators in mycobacteria. Our results show that the terminators that lack a U-trail, variant from the typical *E. coli* intrinsic terminators, are overwhelmingly predominant in all members of the genus. Most terminator structures are concentrated within 50 base pairs downstream of the stop codon. A large number of these terminators occur at the end of experimentally verified or predicted transcription units. We have observed interspecies variations in ΔG and positioning of the terminators downstream of specific genes amongst closely related mycobacterial species suggesting differences in gene expression. The analysis would be useful in furthering our understanding of genome organization and gene expression in mycobacteria, in addition to the improvement in the annotation of the new genomes.

* Corresponding author. Department of Microbiology and Cell Biology, Indian Institute of Science, C.V. Raman Avenue, Bangalore 560012, India. Tel.: +91 80 23600668; fax: +91 80 23602697.
E-mail address: vraj@mcbl.iisc.ernet.in (V. Nagaraja).
[c] Present address: Department of Biochemistry and Bioinformatics, Sri Ramakrishna College of Arts and Science for Women, Coimbatore 641044, India.

1472-9792/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tube.2008.06.004

Introduction

Several members of the genus *Mycobacterium* are deadly pathogens and hence, they continue to be at the center of interest. It is also known that their virulence is a multifaceted phenomenon and involves complex regulatory cascades.[1–4] Efforts to elucidate the molecular mechanisms operating in mycobacteria have shown that the processes such as replication and transcription are qualitatively and quantitatively different from the *E. coli* paradigm.[5] Moreover, the number and usage of sigma factors[6] and the consensus sequences of promoters[7,8] are often dissimilar to that in other well-studied organisms. These features hint at the fact that many aspects of gene regulation are, not unsurprisingly, unique to the genus *Mycobacterium*.

Compared to the transcription initiation process which is currently an active field of investigation in mycobacteria, little is known about termination of transcription in this genus. In *Escherichia coli* and few other bacteria studied, termination is achieved by two mechanisms, employing intrinsic (simple) and factor-dependent (complex) transcription terminators. Functionally, if features of the nascent transcript itself can cause termination in an *in vitro* system, it is called an intrinsic terminator.[9,10] Experiments carried out mainly in *E. coli* have established that a typical intrinsic terminator consists of a GC-rich palindromic region followed by an A-trail on the template strand.[11–14] Transcription of the region results in a RNA which has a hairpin followed by a U-stretch and this has remained the canonical structure for an intrinsic terminator. The sequencing of whole genomes of several eubacteria has resulted in efforts to design algorithms to detect intrinsic terminators across whole genomes. A majority of these algorithms ''search'' for such canonical structures downstream of the stop codon of genes. These algorithms attempt to identify intrinsic terminators based on the strength of the hairpin *and* the weight of the U-trail.[15–18] As a result, bacteria like firmicutes have been shown to be highly dependent on intrinsic terminators.[19] Surprisingly, such analysis resulted in identification of very few intrinsic terminators in several diverse species of bacteria leading to the suggestion that they depend to a very limited degree on intrinsic termination. This anomaly in detecting intrinsic terminators by *in silico* methods was intriguing because intrinsic termination is certainly an economical and efficient process of gene regulation. Moreover, intrinsic terminators from *E. coli* have been shown to function efficiently in other bacteria, suggesting that the basic molecular mechanism is conserved. Hence, it is unclear as to why several groups of bacteria would select against the mechanism.

To address the puzzle and to get a better insight into the prevalence and diversification of transcription terminators, we developed a program called GeSTer (*G*enome *S*canner for *T*erminators) to identify potential intrinsic terminators across whole genomes[20] and analysed the genome of *M. tuberculosis* H37Rv. The results, validated by subsequent *in vitro* and *in vivo* experiments, showed that the majority of functional terminators in *M. tuberculosis* genome lack a U-trail[21]. Sequencing of *M. tuberculosis* genome was followed by that of other mycobacterial genomes as a basis for understanding the genome functions and to decipher the molecular basis of pathogenicity. As a result, several complete mycobacterial genome sequences are known as of now. We have studied the occurrence and organization of intrinsic terminators in 17 *Mycobacterium* genomes, which include 13 different species. Our analysis of the large sample size reveals that mycobacteria depend mainly on intrinsic terminators that lack a U-trail, with considerable inter-species variation. Several of the predicted terminators are located downstream of identified or predicted operons in *M. tuberculosis* H37Rv. The distribution of intrinsic terminators across the entire genome is uniform, and terminators do not show preference for either of the two strands. In addition, most terminators identified seem to have a ΔG in the range of -15 to -25 kcal/mol. The strength and positioning of terminators downstream of specific genes show differences even amongst closely related species indicating variations in gene expression pattern.

Methods

All genomic sequences used in the present study have been downloaded from the genome database of the National Center for Biotechnology Information ftp://ftp.ncbi.nih.gov/genomes/Bacteria/[22] The accepted model of an intrinsic terminator is a double-stranded stem, with a central, unpaired bulb. Besides, there can be symmetric and asymmetric unpaired regions in the stem, termed mismatches and gaps, respectively. For certain terminators, the sequence following the hairpin is important. GeSTer functions by identifying palindromic structures downstream of the genes, and computes their stability, distribution, and the nature of trailing sequences and adjacent structures. GeSTer identifies terminators based on both the ΔG and the permitted lengths of the structural features. The default parameters incorporated into GeSTer are based on the data from the experimentally characterized intrinsic terminators.[23]

The operation of GeSTer is as follows. At first, the algorithm accepts whole genome sequences in the GenBank format and segregates the coding, upstream and downstream regions. Next, it identifies palindromic sequences in -20 to $+270$ nt region around the stop codon, without entering the downstream coding region. Once identified, all possible structures are computed, and the one with the lowest ΔG is retained. The program then reinitiates the search. Lastly, GeSTer uses a minimal $\Delta G_{\text{cut off}}$ filter to select the final set of structures. This filter is based on the genomic GC content of bacteria and characteristics of structures in the upstream region. The basal ΔG of non-coding regions of genomes is species-specific and is strongly correlated to the GC content of the genome. Selection by GeSTer has been optimized so that there is minimum identification of structures that are present upstream of genes. GeSTer then groups the identified terminators based on the sequence content of nucleotides trailing the stem-loop structure, and presence of adjacent structures. Terminators are thus classified into (1) L-shaped/*E. coli* type, which have >3 Us following the stem-loop structure, (2) I-shaped (*Mycobacterium* type), which have <3 Us in the 10 nt stretch following the stem-loop, (3) Tandem/U-shaped,

when two or more structures are present within 50 nt of each other downstream of the same gene, (4) V-shaped, when a stem-loop structure is immediately followed by another with no intervening sequence, and (5) X-shaped/Convergent, which occurs between convergently oriented genes on the 2 strands. The present version of GeSTer is flexible to work with minor variations in GenBank format of input genome sequences. Besides, some details about the genes (such as, gene name) are now available in the output file, making it more user-friendly. Two separate output files have been created which include all the palindromic structures, which were computed but did not meet the genomic $\Delta G_{cut\ off}$. Nevertheless, they could be important in the analysis. The files are called ''weakpalinsreg.dat'' and ''weakpalinscomp.dat''. An auxillary PERL-based program, called TERC (Terminators Represented in Circle) was constructed which plots the positions of terminators on a circular representation of the genome, as given by the GeSTer output. The circular map shows the genome-wide distribution of the two major types of terminators (L-shaped and I-shaped) in both strands of a genome. It also plots the positions of genes based on the coordinates of the start codons. It was designed to observe if there were genomic regions where one kind of terminator is predominant, and also if there are regions where genes have not identified intrinsic terminators. Computational analysis of ΔG of the terminators was carried out with Microsoft Excel 2003. Variation in the position and the strength (ΔG) of the terminators was determined with the GraphPad Prism 4 software using results from GeSTer analysis.

Results and discussion

Occurrence and distribution of intrinsic terminators in *Mycobacterium* genus

We analysed a total of 89,534,161 base pairs from the genomes of 17 *Mycobacterium* genomes, encompassing 13 different species (Supplementary Table 1A–1Q). The complete set of potential terminators (''All'') was compiled for each organism and the most stable structure downstream of each gene was designated to be the ''Best'' terminator. The total number of terminator-like structures identified in mycobacteria is 28,779, of which 21,341 are candidates for ''Best'' structures. These results indicate that, out of a total of 80,887 genes annotated across mycobacteria, 26.4% of them have at least one intrinsic terminator immediately downstream of the coding region (Table 1). This value is on the lower side of the average (best/genes) of 35.5% computed for a sample of 255 bacterial species (Mitra et al., unpublished results). However, the actual number of genes dependent on intrinsic termination would be higher as any given operon contains many genes and is likely to be regulated by a single terminator. These observations could suggest a difference in genome organization in mycobacteria where many more genes may be dependent on a single terminator or on factor-dependent termination.

The overwhelming majority of intrinsic terminators present in all the mycobacterial species analysed are the I-shaped terminators, which have a stem-loop structure, but are devoid of a U-trail (Table 2 and Supplementary Table 1A–Q). In accordance with our previous *in vivo* and *in vitro* studies with *M. tuberculosis* terminators, all the *Mycobacterium* species show a high degree of preference for I-shaped terminators. An average of 89% terminator structures in the species studied are of the I-shape (or *Mycobacterium* type). In contrast, relatively few terminators are of the L-shape (*E. coli* type). It is noteworthy that even in *Mycobacterium leprae* genome, which has comparatively lower GC content (57.8% compared to the *Mycobacterium* average GC content of 66.5%), has 81% I-shaped terminators (Table 2). Our results also explain why other algorithms have failed to detect terminators in several bacterial genomes. By focusing on a hairpin and a stretch of uridylates, even the most recent algorithms have overlooked the variations present in this important group of regulatory structures.[17]

In most *Mycobacterium* species, a number of intrinsic terminators are localized within 50 base pairs after the stop codon (Figure 1, Supplementary Figure 1). This peak in occurrence of terminators is a feature of most bacterial species that we have analysed and it hints at the fact that the spatial distribution of the intrinsic terminators has been optimized such that termination occurs soon after the stop codon. This would be in favour of a more compact genome, typical of prokaryotes, and also prevent unnecessary wastage of ribonucleotides. However, we also note that the distribution of terminator structures after the stop codon is species-specific, even amongst closely related species. For example, *M. tuberculosis*, *Mycobacterium* sp. MCS and *Mycobacterium vanbaalenii* show a similar terminator distribution typical of most eubacteria with a peak around 50 base pairs, that is followed by shoulder peaks. In contrast, *Mycobacterium leprae* shows a much more jagged distribution, while *Mycobacterium ulcerans* has a broad shoulder immediately following the peak. *M. leprae* genome represents a classic example of gene loss. Secondary, yet prominent peaks of terminators are seen in the genome, a pattern also found in few other organisms viz. *Borrelia burgdorferi*, *Treponema pallidum* and *Rickettsia prowazekii*.[20] They all show a modest peak around 50 base pairs with respect to the stop codon and the only feature common to them appears to be that they all are obligate pathogens. More recently, the genomes of three closely related species of mycobacteria known for their ability to degrade organic compounds have been sequenced.[24] All three of them viz. *Mycobacterium* species' JLS, MCS and KMS show an almost identical profile of terminators. GeSTer was also used to identify terminators in the plasmids present in *Mycobacterium gilvum* and *Mycobacterium* species' MCS and KMS. The overall preference for terminators downstream of plasmid-encoded genes mirrors the chromosomal state, with majority being I-shaped terminators (data not shown), indicating the long association between the host and the plasmids.

Intrinsic terminators are uniformly distributed over genomes

The distribution of identified intrinsic terminators in the whole genome has been plotted on a circular map using the algorithm, TERC (Supplementary Figure 2), which uses the coordinates of each terminator identified by GeSTer.

Table 1 Intrinsic terminators in *Mycobacterium* genomes.

Species	Genome (Mb)	Genes	All	Best	Best/genes	$\Delta G_{cut\ off}$
Slow growers						
Mycobacterium avium paratuberculosis	4.83	4397	1782	1250	28.43	−18.2
Mycobacterium avium 104	5.48	5168	2120	1512	29.26	−18.1
Mycobacterium bovis BCG	4.37	4001	1233	936	23.39	−16.9
Mycobacterium leprae TN	3.27	1652	552	445	26.94	−14.3
Mycobacterium tuberculosis H37Rv	4.41	3926	1214	929	23.66	−16.9
Mycobacterium tuberculosis H37Ra	4.42	4081	1275	971	23.79	−16.9
Mycobacterium tuberculosis CDC1551	4.40	4188	1319	1005	24.00	−16.9
Mycobacterium tuberculosis F11	4.42	3988	1260	970	24.32	−16.9
Mycobacterium ulcerans	5.63	4206	1781	1321	31.41	−16.9
Mycobacterium abscessus	5.07	4969	1473	1109	22.32	−16.5
Mycobacterium marinum	6.64	5470	2033	1548	28.30	−17.0
Fast growers						
Mycobacterium smegmatis mc2 155	6.99	6768	2676	1948	28.78	−17.6
Mycobacterium gilvum PYR-GC	5.62	5291	1870	1358	25.67	−17.8
Mycobacterium vanbaalenii	6.49	6033	2128	1578	26.16	−17.7
Mycobacterium sp. JLS	6.05	5792	2062	1515	26.16	−17.9
Mycobacterium sp. MCS	5.71	5444	2000	1473	27.06	−17.9
Mycobacterium sp. KMS	5.74	5513	2001	1473	26.72	−17.9

''All'' denotes total number of structures identified; ''Best'' refers to the strongest structure downstream of the gene and is numerically identical to number of genes with an intrinsic terminator downstream.

For the three representative mycobacterial genomes, the genome-wide predominance of I-shaped terminators is seen. Also, the terminators are uniformly distributed, with no strand bias. There are no visible clusters for the L-shaped terminators. However, there are certain regions where genes do not show any identified terminator. Most of these regions seem to comprise of large operons or, alternatively, rely on Rho-dependent termination.

Table 2 Frequency of occurrence of different types of terminators in genus *Mycobacterium*.

Species	Terminator types*				
	%L†	%I†	%V‡	%U‡	%X‡
Slow growers					
Mycobacterium avium paratuberculosis	7.04	92.96	0.17	12.07	2.24
Mycobacterium avium 104	7.80	92.20	0.28	11.42	2.31
Mycobacterium bovis BCG	8.55	91.45	0.41	8.43	2.11
Mycobacterium leprae TN	18.88	81.12	0.00	6.88	0.36
Mycobacterium tuberculosis H37Rv	8.93	91.07	0.41	8.48	1.98
Mycobacterium tuberculosis H37Ra	8.44	91.56	0.39	8.24	1.96
Mycobacterium tuberculosis CDC1551	9.35	90.65	0.30	8.26	1.97
Mycobacterium tuberculosis F11	8.97	91.03	0.48	8.33	1.98
Mycobacterium ulcerans	11.28	88.72	0.11	9.94	2.02
Mycobacterium abscessus	14.78	85.21	0.2	8.89	3.46
Mycobacterium marinum	10.79	89.21	0.1	9.05	2.71
Fast growers					
Mycobacterium smegmatis mc2 155	13.45	86.55	0.07	9.94	3.81
Mycobacterium gilvum PYR-GC	9.87	90.13	0.27	9.41	1.87
Mycobacterium vanbaalenii	9.89	90.11	0.19	8.74	2.63
Mycobacterium sp. JLS	8.38	91.62	0.15	10.28	2.62
Mycobacterium sp. MCS	9.10	90.90	0.25	9.90	2.65
Mycobacterium sp. KMS	9.37	90.63	0.25	9.85	2.65

*(1) L-shaped, which have >3 Us following the stem-loop structure, (2) I-shaped, which have <3 Us in the 10nt stretch following the stem-loop, (3) Tandem/U-shaped, when 2 or more structures are present within 50 nt of each other downstream of the same gene, (4) V-shaped, when a stem-loop structure is immediately followed by another with no intervening sequence, and (5) X-shaped/Convergent, which occur between convergently oriented genes on the 2 strands.

† L-shaped and I-shaped terminators are expressed as percentage of ''Best'' structures.

‡ U, V and X structures are expressed as percentage of ''All'' structures.

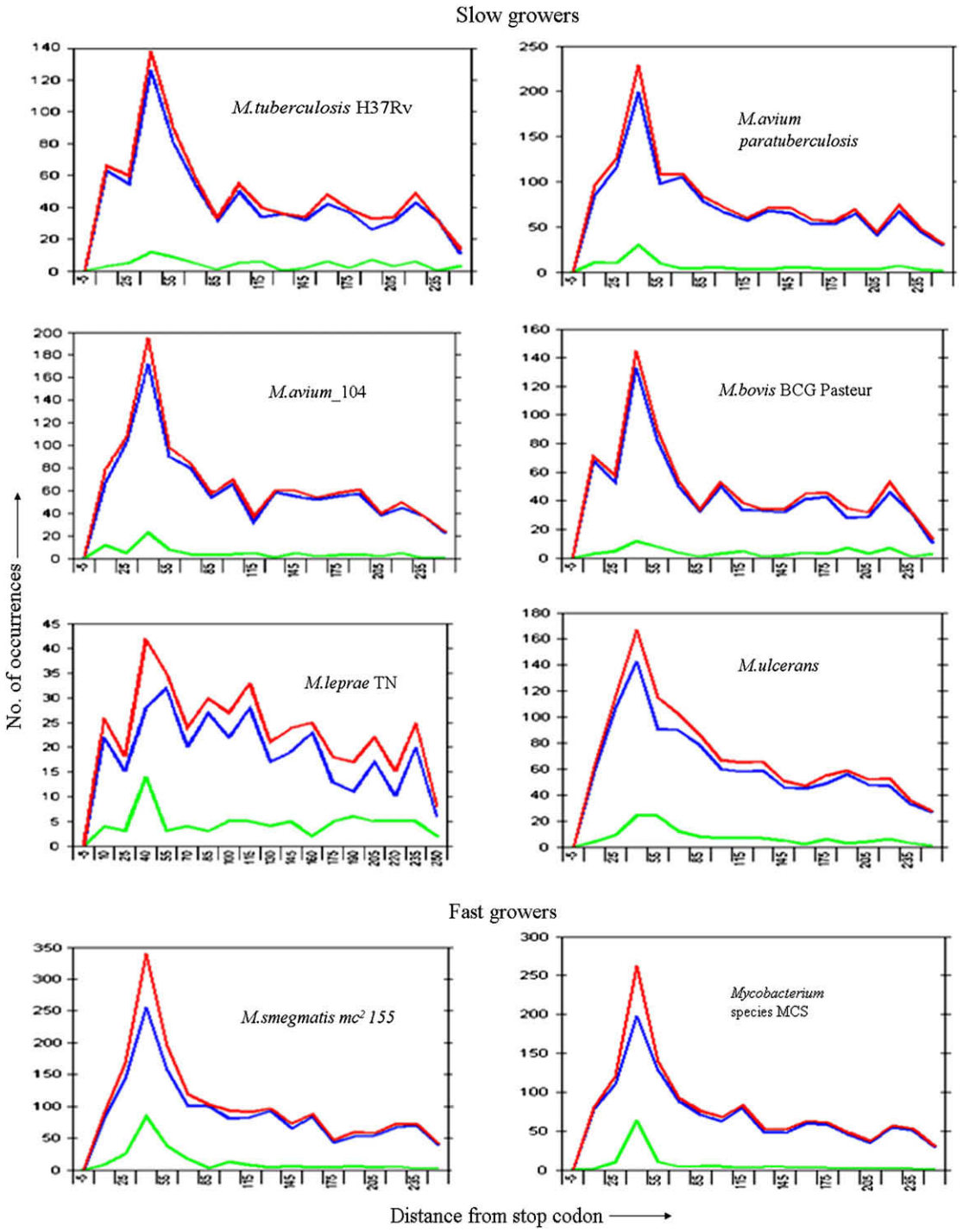

Figure 1 Distribution of L-shaped and I-shaped terminators in *Mycobacterium* species. The abscissa denotes the distance of terminator from stop codon, while the ordinate denotes number of occurrences. Red, blue and green denote distribution of Best, I-shaped and L-shaped structures. respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Most identified terminators have an ''optimized'' stem-length and ΔG

One of the primary determinants of termination efficiency by an intrinsic terminator is the strength or stability of the stem-loop structure (as given by its ΔG). A stronger structure (more negative ΔG value) will probably cause more efficient pausing and termination, while a weaker structure (less negative ΔG value) is likely to have increased read through by the RNA polymerase juggernaut. As described in Methods GeSTer employs a genome-specific $\Delta G_{cut\ off}$ (Table 1). Structures which have a ΔG more

negative than the $\Delta G_{cut\ off}$ value are only considered to be functional terminators. When the terminators were sorted as per their strength, it was seen that in all *Mycobacterium* species, 60–70% of the identified terminators have ΔG values only slightly more negative than the $\Delta G_{cut\ off}$. Across mycobacteria, majority of the identified ''best'' structures have ΔG in the range of (−15 to −25) kcal/mol (Figure 2). Thus, there seems to be an optimization of terminator strength as shown previously with canonical intrinsic terminators. For e.g., in case of the λtR2 terminator and it's mutants, the optimal hairpins which maximize termination efficiency have a stem-length of 7–9 base pairs and such terminators have a ΔG in the range of −15 to −25 kcal/mol.[14] Also, our results with mycobacteria are in agreement with the average ΔG range found for terminators from both *E. coli* (−14 kcal/mol) and *Bacillus subtilis* (−16 kcal/mol) genomes.[19,23] Thus, functional optimization of the terminator structure and strength seems to be a general feature. In our earlier experimental verification carried out to validate GeSTer, we used only the strongest terminators having ΔG values around −50 kcal/mol. While these very strong terminators (*tuf* and *Rv1324)* were highly efficient,[21] the present results show that, as regards ΔG, terminators with very negative ΔG are rare in mycobacteria. Now, to confirm the occurrence of relatively weak or ''functionally optimized'' terminators, we have resorted to analyse the experimental data of gene expression as also of predicted operon maps. The results for *M. tuberculosis* H37Rv (Table 3) indicate that GeSTer-determined terminators are located downstream of many identified and predicted operons and transcription units. Moreover, in most of these cases, the terminator has a ΔG value in the range of −16 to −25, hence validating GeSTer output.

It is possible that several factors, such as NusA, contribute to increase the efficiency of these terminators *in vivo*. The majority of terminators identified here have a stem-length of 7–17 base pairs (data not shown). Since bacterial genomes need to be densely packed, it would be advantageous for cells to evolve intrinsic terminators, which are encoded by smaller genomic regions, but nevertheless are functionally efficient. On the other hand, the weak terminators could be more easily subjected to regulation which could account for their prevalence in the genomes.

GeSTer-identified terminators occur at the ends of many operons/transcription units in *Mycobacterium* genomes

To validate the GeSTer predictions, we analysed the experimental data obtained for *M. tuberculosis* H37Rv.[25] This comprehensive dataset of operon map relies on both microarrays and RT-PCR, as well as bioinformatics approaches. In the representative data tabulated (Table 3, Supplementary Table 2), we find a significant number of operons/transcription units have an intrinsic terminator at the 3′ end. Intrinsic terminators are found in operons that encode both housekeeping genes as well as genes implicated in virulence. Examples include the induced *iniBAC* operon, the PPE operon of *Rv2431c-Rv2430c*, *Rv3083-Rv3089* operon (encoding *fad* gene), the *senX3-regX3* operon that codes for a two-component sensory system and the *esxB-esxA* operon (encoding CFP-10 and ESAT-6). Besides, as shown in the tables, many terminators located downstream of operons have a ΔG value between −16 and −24 kcal/mol. This provides evidence that such ''weak'' terminators, which are abundant in mycobacteria and also in other genomes analysed are indeed functional terminators that are effective in countering the advance of the RNA polymerase juggernaut.

The rRNA operons are organized in the order of 16S rDNA–23S rDNA–5S rDNA in all bacteria. To extend our validation of GeSTer-predicted terminators to other mycobacterial genomes, we analysed the presence of terminators downstream of the 5S rRNA-encoding gene in mycobacterial genomes. The results (Table 4) show the presence of the terminators at the 3'end for these operons. This data thus extends the reliability of GeSTer to predict one ''boundary'' of operons across genus *Mycobacterium*.

Differential strength and positioning of terminators downstream of a specific gene

The basis to the species-specific differences in whole-genome terminator profiles was analysed by comparing the spatial location and strength (ΔG) of intrinsic terminators present downstream of specific genes across the different *Mycobacterium* species. Representative genes involved in ''housekeeping'' function were chosen for this purpose as the gene products would have important roles in metabolism and gene expression (Figure 3). The results show that there is considerable variation in the positioning of terminator with respect to the stop codon downstream of a specific gene and it's strength/stability, as denoted by the ΔG value of the stem-loop structure (Figure 3). For example, the *tuf* (gene encodes EF-Tu) terminator initiates at position +14 for *M. tuberculosis* H37Rv and at +45 for *M. leprae*. For *rpoC* (β′ subunit of RNAP), the downstream terminator starts at nucleotide position −3 in case of *Mycobacterium ulcerans*, at +12 for *Mycobacterium smegmatis* and at +216 for *M. leprae*. In case of the *topA* (topoisomerase I), the ''best'' terminator is located between 150–200 base pairs downstream of the stop codon for 5 species, while it is found at +58 for *M. ulcerans*. A similar positional variation is observed for the terminator downstream of *thrB* (threonine kinase) gene. The stability of the stem-loop structure also varies for a given terminator across mycobacterial species (Figure 3). Amongst many examples, this is well illustrated in case of *tuf* terminators from *M. tuberculosis* H37Rv and *M. leprae* (ΔG of −49.6 and −19.8 kJ/mol, respectively). Moreover, we find that a specific terminator, while present in most species, is absent in the corresponding position in few other species. For e.g., there is no identifiable terminator downstream of the *topA* gene in *M. smegmatis* and *M. leprae*, and no terminator is picked up by GeSTer downstream of the *thrB* (threonine kinase) gene in *M. avium* and *M. vanbaalenii*. For all the genes considered, the overall gene neighbourhood and also the distance from the downstream gene is similar across the genus *Mycobacterium*, minimizing the possibility that the positional variation of terminators observed is a function of different gene neighbourhoods.

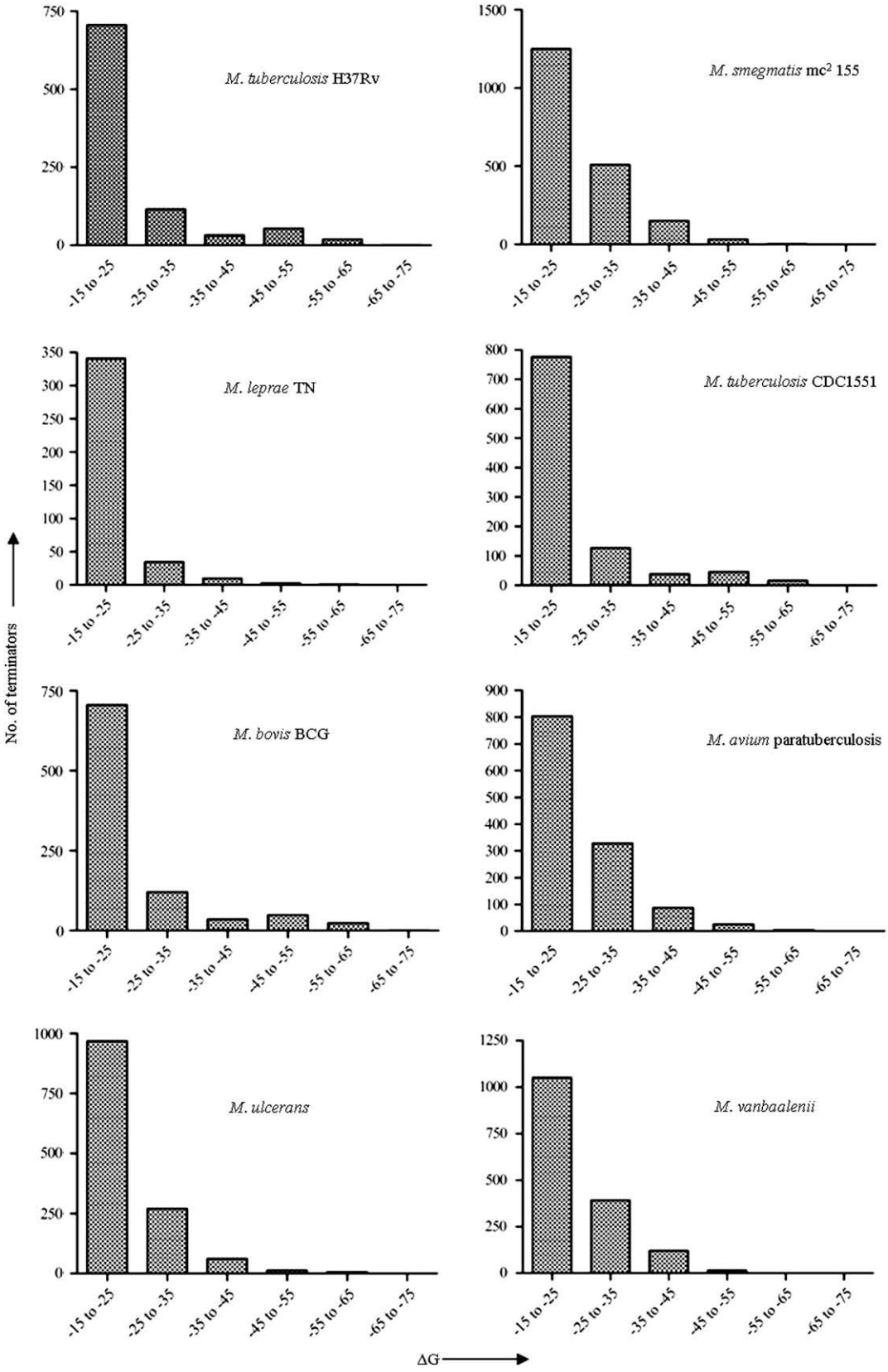

Figure 2 Distribution of identified intrinsic terminators as per their ΔG in *Mycobacterium* species. The abscissa denotes ΔG. The ordinate denotes number of terminators.

Table 3 Intrinsic terminators downstream of operons/genes in *M. tuberculosis* H37Rv*.

Gene/operon	Description	ΔG of terminator
iniBAC (Rv0341-0343)	Three genes in tandem induced by isoniazid	−21
Rv0967-0968-ctpV-0970	Copper sensitive operon (cso)	−24
Rv3083-84-85-86-87-88-89	*mymA* operon	−18.1
Rv2431c-Rv2430c	PE-PPE operon	−21
Rv0667-Rv0668	*rpoB-rpoC* (β, β′ subunits of RNAP)	−52
Rv1303-Rv1312	Operon encoding F0F1 ATPase subunits	−18.7
Rv0072-0073	Probable Gln-transporter subunits; predicted operon	−18.3
Rv0490-Rv0491	SenX3-RegX3 (2-component sensory system)	−26
trpG (or pabA/Rv0013)	*p*-Aminobenzoate synthase component II	−20.6
Rv3874-Rv3875	*esxA-esxB* (encoding ESAT-6) operon	−33
Rv3863	Gene from RD1 locus	−17.1
Rv3878	Gene from RD1 locus	−17.4

* partial list.

Conclusions

Intrinsic termination is an economical and efficient means for spatial regulation of gene expression. At a whole-genome level, terminators function to compartmentalize gene expression. Most evidence indicates that the hairpin causes a pausing of the RNA polymerase. It also weakens the interactions between RNA polymerase, and the DNA—RNA hybrid formed between transcript and the template. The weak nature of dA: rU base pairing is considered to facilitate the release of the transcript.[13,14,26] Two different models — ''forward translocation model''[27—29] and ''allosteric model''[30] provide alternate views on the mechanism of intrinsic termination.

Our results provide a detailed account of the occurrence and distribution of terminators in the genus *Mycobacterium* and indicate a variant structure lacking ''U'' trail is predominant. It is most likely that structural variants of intrinsic terminators existed in the ancestral prokaryotes. Subsequent evolutionary history resulted in selection of different terminator types in different species. Since L-shaped terminators would have become rare in a GC-rich genome, I-shaped terminators were selected as the major type. Genomic GC content is, however, unlikely to be the primary determinant of terminator type as it is known that bacteria with similar GC contents have widely different preference for I-shaped terminators.[20] It must be noted that the function of the U-trail, has remained ambiguous. Although it is considered crucial for efficient termination,[31] several terminators have been shown to retain termination efficiency in absence of the U-trail or when many of the U residues have been deleted. Hence, it seems, unlike the hairpin structure, the U-trail can be dispensable at certain intrinsic terminators. Indeed, in the case of mycobacterial species, the U-trail could have become functionally redundant in a situation where the RNA polymerase is transcribing at a slower rate compared to other bacteria, such as *E. coli*.[5] This seems plausible as the function attributed to the U-trail is to enhance the stalling of the elongating RNA polymerase. Sequences downstream of the termination site can functionally complement the U-trail's role in causing termination. It has been recently shown that GC-rich sequences immediately downstream of an intrinsic terminator cause more efficient termination than AT-rich structures.[30] In GC-rich genomes such as those of *Mycobacterium,* there is a higher probability of a GC-rich region downstream of a terminator, and this could aid in termination. Besides, lineage specific differences in the domainal organization of β and β′ subunits of RNA polymerase have been reported.[32] Thus, it appears that the mycobacterial transcription machinery could have co-evolved to terminate more efficiently at I-shaped terminators.

Intrinsic terminators, being non-protein coding structures, are subject to lesser sequence constraints. This would have facilitated the selection in favour of optimal strength and positioning of intrinsic terminators, such as to suit the needs of a particular organism, yet retaining the common molecular mechanism of the termination process. A consensus length and ΔG of the hairpin stem of terminators is observed, which is in agreement with experimental[14]

Table 4 Intrinsic terminators downstream of *rrn* operons in different *Mycobacterium* species.

Species	*rrn* Operon (s)	Identified terminator	ΔG*	ΔG†
Slow growers				
Mycobacterium avium paratuberculosis	1	Yes	−24.7	
Mycobacterium bovis BCG	1	Yes	−23.7	
Mycobacterium tuberculosis H37Rv	1	Yes	−23.7	
Mycobacterium abscessus	1	Yes	−247	
Fast growers				
Mycobacterium smegmatis mc2 155	2	Yes	−21.3	−30.2
Mycobacterium gilvum PYR-GC	2	Yes	−29.5	−27.5
Mycobacterium vanbaalenii	2	Yes	−29.2	−28
Mycobacterium sp. MCS	2	Yes	−33.1	−33.1

*, † For genomes where there are two rrn operons, the ΔG has been separately tabulated for each operon.

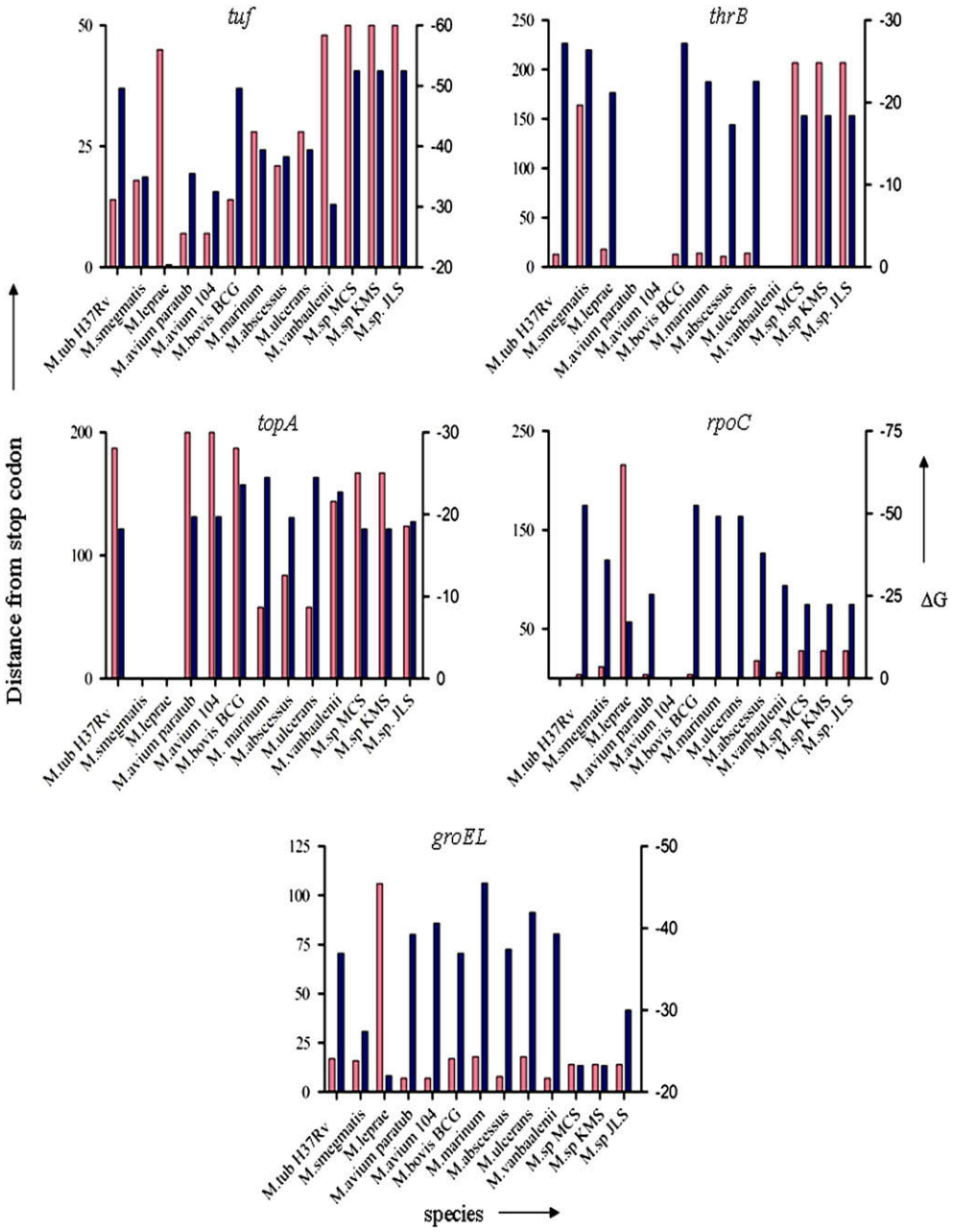

Figure 3 Comparison of ΔG and distance from stop codon of terminators downstream of specific genes across several *Mycobacterium* species. The terminators shown here are directly downstream of important genes, *rpoC* (β′ subunit of RNA Pol), *thrB* (threonine kinase), *topA* (topoisomerase I), *tuf* (elongation factor, EF-Tu), *groEL* (chaperonin GroEL). The left *Y*-axis denotes distance of terminator from stop codon (pink); the right *Y*-axis denotes ΔG of the terminator (blue). Absence of bars signifies that there is no intrinsic terminator in case of that particular species. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

and bioinformatics analysis.[19,23] Terminators predicted by GeSTer occur immediately downstream of many predicted and experimentally observed operons in *M. tuberculosis*. Thus, the genes and operons, irrespective of their function, would utilize this efficient regulatory and anti-read through mechanism. With many functionally unidentified genes present in mycobacterial species' and a number of genomes yet to be sequenced, our analysis could be important in improving genome annotation, and operon prediction. GeSTer would be an important tool in understanding genome organization and regulation of gene expression across genus *Mycobacterium*.

Acknowledgements

We thank Ranjana Prakash and O. Krishnadev for algorithm development and analysis. The work is supported by grants from Indo-Swiss Joint Research Programme/Department of Science and Technology, Government of India and Center

for Excellence in mycobacteria research by Department of Biotechnology, Government of India.

Funding: See above.

Competing interests: None declared.

Ethical approval: Not required.

Supplementary material

Supplementary material can be found, in the online version, at doi:10.1016/j.tube.2008.06.004.

References

1. Clark-Curtiss JE, Haydel SE. Molecular genetics of *Mycobacterium tuberculosis* pathogenesis. *Annu Rev Microbiol* 2003;**57**:517—49.
2. Brosch R, Pym AS, Gordon SV, Cole ST. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* 2001;**9**:452—8.
3. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *J Bacteriol* 1996;**178**:1274—82.
4. Collins DM. Virulence factors of *Mycobacterium bovis*. *Tuberculosis (Edinb)* 2001;**81**:97—102.
5. Harshey RM, Ramakrishnan T. Rate of ribonucleic acid chain growth in *Mycobacterium tuberculosis* H37Rv. *J Bacteriol* 1977;**129**:616—22.
6. Manganelli R, Provvedi R, Rodrigue S, Beaucher J, Gaudreau L, Smith I. Sigma factors and global gene regulation in *Mycobacterium tuberculosis*. *J Bacteriol* 2004;**186**:895—902.
7. Unniraman S, Chatterji M, Nagaraja V. DNA gyrase genes in *Mycobacterium tuberculosis*: a single operon driven by multiple promoters. *J Bacteriol* 2002;**184**:5449—56.
8. Smith I, Bishai WR, Nagaraja V. Control of mycobacterial transcription. In: Cole ST, editor. *Tuberculosis and the Tubercle bacillus*. Washington, D.C: ASM Press; 2005. p. 219—31.
9. von Hippel PH. An integrated model of the transcription complex in elongation, termination, and editing. *Science* 1998;**281**:660—5.
10. Richardson JP, Greenblatt J. Control of RNA chain elongation and termination. In: Neidhert FC, Curtiss III R, Ingraham JL, Lin ECC, Low KB, Magasanik B, et al., editors. *Escherichia coli and Salmonella: cellular and molecular biology* 2nd ed. Washington, DC: American Society for Microbiology; 1996. p. 822—48.
11. Yarnell WS, Roberts JW. Mechanism of intrinsic transcription termination and antitermination. *Science* 1999;**284**:611—5.
12. Farnham PJ, Platt T. Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro*. *Nucleic Acids Res* 1981;**9**:563—77.
13. Artsimovitch I, Landick R. Interaction of a nascent RNA structure with RNA polymerase is required for hairpin-dependent transcriptional pausing but not for transcript release. *Genes Dev* 1998;**12**:3110—22.
14. Wilson KS, von Hippel PH. Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc Natl Acad Sci U S A* 1995;**92**:8793—7.
15. Ermolaeva MD, Khalak HG, White O, Smith HO, Salzberg SL. Prediction of transcription terminators in bacterial genomes. *J Mol Biol* 2000;**301**:27—33.
16. Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ. Prediction of rho-independent transcriptional terminators in *Escherichia coli*. *Nucleic Acids Res* 2001;**29**:3583—94.
17. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 2007;**8**:R22.
18. Washio T, Sasayama J, Tomita M. Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Res* 1998;**26**:5456—63.
19. de Hoon MJ, Makita Y, Nakai K, Miyano S. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comput Biol* 2005;**1**:e25.
20. Unniraman S, Prakash R, Nagaraja V. Conserved economics of transcription termination in eubacteria. *Nucleic Acids Res* 2002;**30**:675—84.
21. Unniraman S, Prakash R, Nagaraja V. Alternate paradigm for intrinsic transcription termination in eubacteria. *J Biol Chem* 2001;**276**:41850—5.
22. NCBI GenBank FTP Site. Available from: <ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria>.
23. d'Aubenton Carafa Y, 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—58.
24. Miller CD, Hall K, Liang YN, Nieman K, Sorensen D, Issa B, et al. Isolation and characterization of polycyclic aromatic hydrocarbon-degrading *Mycobacterium* isolates from soil. *Microb Ecol* 2004;**48**:230—8.
25. Roback P, Beard J, Baumann D, Gille C, Henry K, Krohn S, et al. A predicted operon map for *Mycobacterium tuberculosis*. *Nucleic Acids Res* 2007;**35**:5085—95.
26. Gusarov I, Nudler E. The mechanism of intrinsic transcription termination. *Mol Cell* 1999;**3**:495—504.
27. Santangelo TJ, Roberts JW. Forward translocation is the natural pathway of RNA release at an intrinsic terminator. *Mol Cell* 2004;**14**:117—26.
28. Lynn SP, Kasper LM, Gardner JF. Contributions of RNA secondary structure and length of the thymidine tract to transcription termination at the thr operon attenuator. *J Biol Chem* 1988;**263**:472—9.
29. Abe H, Aiba H. Differential contributions of two elements of rho-independent terminator to transcription termination and mRNA stabilization. *Biochimie* 1996;**78**:1035—42.
30. Epshtein V, Cardinale CJ, Ruckenstein AE, Borukhov S, Nudler E. An allosteric path to transcription termination. *Mol Cell* 2007;**28**:991—1001.
31. Ryder AM, Roberts JW. Role of the non-template strand of the elongation bubble in intrinsic transcription termination. *J Mol Biol* 2003;**334**:205—13.
32. Iyer LM, Koonin EV, Aravind L. Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. *Gene* 2004;**335**:73—88.