Microbiology (2015), **161**, 719–728 DOI 10.1099/mic.0.000037

Transcription and translation of the *rpsJ*, *rplN* and rRNA operons of the tubercle bacillus

Teresa Cortes and Robert Ashley Cox

Division of Mycobacterial Research and Division of Mathematical Biology, MRC National Institute for Medical Research, Mill Hill, London, NW7 1AA, UK

Correspondence
Robert Ashley Cox
rcox@nimr.mrc.ac.uk

Received 19 September 2014
Accepted 18 January 2015

Several species of the genus *Mycobacterium* are human pathogens, notably the tubercle bacillus (*Mycobacterium tuberculosis*). The rate of proliferation of a bacterium is reflected in the rate of ribosome synthesis. This report describes a quantitative analysis of the early stages of the synthesis of ribosomes of *M. tuberculosis*. Specifically, the roles of three large operons, namely: the *rrn* operon (1.7 microns) encoding *rrs* (16S rRNA), *rrl* (23S rRNA) and *rrf* (5S rRNA); the *rpsJ* operon (1.93 microns), which encodes 11 ribosomal proteins; and the *rplN* operon (1.45 microns), which encodes 10 ribosomal proteins. A mathematical framework based on properties of population-average cells was developed to identify the number of transcripts of the *rpsJ* and *rplN* operons needed to maintain exponential growth. The values obtained were supported by RNaseq data. The motif 5′-gcagac-3′ was found close to 5′ end of transcripts of mycobacterial *rplN* operons, suggesting it may form part of the RpsH feedback binding site because the same motif is present in the ribosome within the region of *rrs* that forms the binding site for RpsH.

INTRODUCTION

The genus *Mycobacterium* is of interest because several species are pathogenic to man and the diseases they cause are difficult to treat. For example, in humans tuberculosis is caused by *Mycobacterium tuberculosis*, leprosy by *Mycobacterium leprae*, and *Mycobacterium abscessus* causes a lung disease among patients suffering from cystic fibrosis (Ripoll *et al.*, 2009). Furthermore, tuberculosis in cattle is caused by *Mycobacterium bovis* and *Mycobacterium marinum* is pathogenic to fish (Tobin & Ramakrishnan, 2008). *M. bovis* BCG is non-pathogenic and is often used to provide a frame of reference for the pathogens, and so is *Mycobacterium smegmatis*. It is desirable to understand how mycobacteria proliferate in order to improve the medical treatments available and to counteract the threats of drug resistance.

The foundations for studying the properties of bacterial cell cultures were laid in the period 1958–1976. First, Schaechter *et al.* (1958) studied the relative masses of DNA, RNA or protein (DNA : RNA : protein) of bacterial cultures grown in different media. They showed that bacterial cells increased in size with increasing growth rate and that the number of ribosomes increased to meet the demand for a faster rate of protein synthesis. They also showed that the notion of an average cell is conceptual because it represents cells of all ages from the newly born (age a=0) to cells about to divide (age a=1). Because the average cell reflects the entire age distribution it is more accurately termed a population-average cell, which we refer to here as a 'cell' for brevity.

Second, a number of studies (Byrne *et al.*, 1964; Miller *et al.*, 1970; Stent, 1964) showed that gene transcripts were translated as they were transcribed (transcription/translation coupling). Third, the exponential growth equation provides the basis for a mathematical approach for bacterial cell growth. Fourth, knowledge of both the ratios DNA : RNA : protein and the size of the genome allows the macromolecular composition of an average cell to be estimated (Cox, 2004). The work of Miller *et al.* (1970) provided the first view of a large functional operon (Fig. 1) and also revealed that large operons are uncommon.

Ribosome synthesis is essential to cell proliferation. The importance of ribosomes is reflected in the ways that the required genes are organized within the genome; notably, because a high proportion are located within three large operons, namely: the rRNA operon (5167 bp, 1.76 microns) encoding the three components *rrs*, *rrl* and *rrf*, which are 16S rRNA, 23S rRNA and 5S rRNA, respectively; the *rpsJ* operon (5679 bp, 1.93 microns), which comprises 11 ribosomal protein genes; and the *rplN* operon (4257 bp, 1.45 microns), which comprises 10 ribosomal protein genes. The study of these operons is facilitated by the wealth of information

Abbreviation: RNAP, RNA polymerase.

The Array Express accession number for the sequence data reported in this paper is E-MTAB-3252.

Three supplementary tables and two supplementary figures are available with the online Supplementary Material.

Fig. 1. Historic electron micrograph of the transcription/translation of an operon of *Escherichia coli* (μ=0.69 h^{-1}). The section of DNA being transcribed corresponds to the *rplN* operon in length. Image reproduced from Miller & Hamkalo (1972) with permission, licence number 3525270508092.

available for the structure and function of ribosomes and ribosomal components (Bashan & Yonath, 2008; Williamson, 2009).

The structure and expression of the rRNA (*rrn*) operon of *M. tuberculosis* has been studied (Kempsell *et al.*, 1992) and compared with *rrn* operons of other mycobacteria (Gonzalez-y-Merchand *et al.*, 1996, 1997; Ji *et al.*, 1994a, b, c). In contrast, no studies of the two mycobacterial operons *rpsJ* and *rplN* that together encode 21 ribosomal proteins have been reported. For this reason, we have studied the properties of these two operons with the aim of defining key steps in the proliferation of the tubercle bacillus.

Although the *rpsJ* and *rplN* operons of mycobacteria resemble those of *Escherichia coli* in structure, the ways by which they are regulated may differ because of the large difference in the maximum growth rates of the two species.

Electron micrographs of these operons undergoing transcription and translation were obtained by Miller *et al.* (1970) after lysis of cells of *E. coli* grown with a generation time of 1 h (Fig. 1). The number of transcripts reveals the activity of the operon. In principle, this figure provides a reference point for the study of the mycobacterial homologues. The required mathematical framework was developed and used to quantify the expression of the mycobacterial operons. Both the sequence and the protein co-factors needed for function were identified.

METHODS

Culture conditions and RNA isolation. *M. tuberculosis* H37Rv was grown in Middlebrook 7H9 medium supplemented with 0.2 % w/v glycerol and 10 % Albumin Dextrose Catalase (ADC) supplement in roller bottle culture. RNA was isolated from mid-exponentially growing bacteria as described by Arnvig *et al.* (2011). RNA was treated with Turbo DNase (Ambion) until DNA free. The quality of RNA was assessed using a Nanodrop spectrophotometer (ND-1000; Labtech) and an Agilent bioanalyser.

RNA sequencing. A single RNA sample was used by vertis Biotechnologie to construct a cDNA library for whole transcriptome analysis. RNA was fragmented with ultrasound (four pulses of 30 s at 4 °C), then treated with Antarctic phosphatase and rephosphorylated with polynucleotide kinase (PNK). Afterwards, the fragmented RNA was poly(A)-tailed using poly(A) polymerase and the Illumina 5′TruSeq adaptor was ligated to the 5′-phosphate of the RNA. First-strand cDNA synthesis was performed using an oligo(dT)-adaptor primer and M-MLV reverse transcriptase. The resulting cDNA was amplified by PCR. The cDNA library was sequenced as single-end reads on an Illumina HiSeq 2000 system.

Read mapping. FastQC (Babraham Institute) was used for quality control of the Illumina produced FASTQ file. Poor-quality read bases were trimmed using the SolexaQA package (Cox *et al.*, 2010); default parameters were used, trimming bases with confidences $P>0.05$, and removing reads <25 bases. Good-quality reads were mapped to the reference sequence of *M. tuberculosis* H37Rv (GenBank/EMBL/DDBJ accession no. AL123456) as single end data using Burrows–Wheeler Aligner (BWA) (Li & Durbin, 2009). Transcriptome coverage, defined as the number of reads mapped per bp of the H37Rv genome, was calculated using BEDTools (Quinlan & Hall, 2010) and was found to be 432-fold. The number of reads mapped to each annotated ORF was calculated using BEDTools ranging from 1 to 19 953. Calculation of pairwise correlation coefficients demonstrated a high degree of reproducibility between this dataset (Spearman *r* ranging from 0.80 to 0.85; see Fig. S1, available in the online Supplementary Material) and previously published transcriptome analyses of exponentially growing *M. tuberculosis* by RNaseq (Arnvig *et al.*, 2011; Cortes *et al.*, 2013).

Mathematical equations needed for quantitative analysis of transcription and translation. During exponential growth a cell component x such as RNA or protein will increase with time according to equation (A1) where x is the amount at time t, t_0 is the time at a reference point and μ is the specific growth rate (h^{-1}).

$$x = x_0 \bullet e^{\mu t} \tag{A1}$$

The specific synthesis rate, ω_x, at which x increases is given by the equalities in equation (A2)

$$\varpi_x = \mathrm{d}x/\mathrm{d}t = \mu \bullet x_0 \bullet e^{\mu t} = \mu \bullet x \tag{A2}$$

The rate $\omega_{c-p(i)}$ at which the number $n_{c-p(i)}$ of copies of protein $p_{(i)}$ increases is given in equation (A3).

$$\varpi_{c-p(i)} = \mu \bullet n_{c-p(i)} \tag{A3}$$

The specific synthesis rate is also the product of the number $n_{R(i)}$ of ribosomes synthesizing $p_{(i)}$ and the rate $\varepsilon_{\mathrm{aa(i)}}$ amino acids h^{-1} at which amino acids are incorporated into nascent polypeptide chain. Equating the right hand sides of equations (A2) and (A3) leads to equation (A4)

$$\mu \bullet n_{c-p(i)} \bullet l_{aa(i)} = n_{R(i)} \bullet \varepsilon_{aa(i)} \tag{A4}$$

where $l_{aa(i)}$ amino acids is the length of $p_{(i)}$. Rearranging equation (A4) to make $n_{R(i)}$ the subject leads to equation (A5).

$$n_{R(i)} = n_{c-p(i)} \bullet l_{aa(i)} \bullet (\mu / \varepsilon_{aa(i)}) \tag{A5}$$

A conversion factor $n_{R(i)/tr(i)}$ is needed to relate the number of ribosomes to a transcript of $\mathrm{ORF_{(i)}}$ encoding $p_{(i)}$ [see equation (A6)].

$$n_{tr(i)} \bullet n_{R(i)/tr(i)} = n_{c-p(i)} \bullet l_{aa(i)} \bullet (\mu / \varepsilon_{aa(i)}) \tag{A6}$$

As a consequence of transcription/translation coupling it is thought that the moving ribosome controls the rate of transcription by preventing RNA polymerase (RNAP) from spontaneously backtracking (Burmann *et al.*, 2010; Proshkin *et al.*, 2010). Consequently, both transcription and translation are determined by codon usage and the availability of nutrients. We consider, as a first approximation, that the peptide chain elongation rate, $\varepsilon_{aa(\mathrm{av})}$, derived from the ratios DNA : RNA : protein is representative of the rate, $\varepsilon_{aa(i)}$, for $\mathrm{ORF}_{(i)}$.

RESULTS AND DISCUSSION

Inspection of the relevant mycobacterial genomic data reveals the presence of three large operons that encode proteins: one encoding components of ATP synthase and two operons, namely, *rpsJ* and *rplN*, that encode ribosomal proteins. The *rpsJ* and *rplN* operons of *E. coli* have been studied (Mattheakis *et al.*, 1989; Stelzl *et al.*, 2003). In general, the *rpsJ* and *rplN* operons are found among a cluster of ribosomal proteins (Coenye & Vandamme, 2005) modified by deletions and by the introduction of non-ribosomal protein genes through horizontal gene transfer. There are no reports of detailed studies of these operons in mycobacteria.

The principal features of the above-mentioned operons are evident from Fig. 1: namely, the length of the DNA segment, the number of transcripts and the number of ribosomes per transcript. The number of transcripts per operon would be expected to depend on growth rate. Mycobacteria grow slowly compared with *E. coli* and so we would expect fewer transcripts per operon.

Number of transcripts of a gene present per cell

Number of transcripts calculated from the growth equation. The mathematical treatment (see Methods) is based on the exponential growth equation that defines the relation between the specific growth rate, μ;, the number of copies $n_{c-p(i)}$ of a protein $p_{(i)}$ of length $l_{aa(i)}$ amino acids encoded by $ORF_{(i)}$ and the number of transcripts $n_{tr(i)}$ needed to maintain the rate of synthesis that is required: see equation (1) where $n_{R(i)/tr(i)}$ is the number of ribosomes per transcript [see also equation (A6) of Methods].

$$n_{tr(i)} \bullet n_{R(i)/tr(i)} = n_{c-p(i)} \bullet l_{aa(i)} \bullet (\mu / \varepsilon_{aa(i)}) \qquad (1)$$

The number of transcripts of a gene was obtained by evaluating the parameters of the right hand side of equation (1), which is the required form of the growth equation. Three assumptions were made in applying this equation. First, it was assumed that, except for *rplL*, ribosomal proteins are present as one copy per ribosome. Second, ribosomal proteins were assumed to be located mainly in ribosomes; in other words, the number of copies of a ribosomal protein was assumed to be approximately equal to the product of the number of ribosomes and the number of copies of that protein per ribosome. Third, it was assumed that the peptide chain elongation rate was approximately equal to the polypeptide chain elongation rate of the protein fraction of the population-average cell. However, the rate of synthesis of individual proteins may vary according to codon usage.

Evaluation of the number of transcripts per gene from RNaseq data. The first step was to relate the numbers of partially sequenced gene transcripts revealed by the RNaseq platform to the numbers of gene transcripts of ORFs encoding both RNA and protein components of the ribosome. The principal steps in the RNaseq procedure starting from a sample (n_{cells}) of a bacterial culture and ending with the number [$n_{p\text{-}tr(i)}$] of partially sequenced transcripts assigned to an ORF [$\mathrm{ORF}_{(i)}$] are shown in Methods. The crucial factor is that only a section of 25–70 terminal nucleotides of each mRNA fragment is measured. The limitation of this approach is shown (Fig. 2) by the data for each of the three rRNA components *rrs, rrl* and *rrf*. The profiles found for the number of gene transcripts at each point from the 5′ to the 3′ end of the gene are shown for the mature species in Fig. 2(a) and the profiles revealed by the RNaseq platform are shown in Fig. 2(b). If the sequences of the RNA fragments produced by the RNaseq procedure had been complete the two profiles would have been identical within experimental error. The partial sequencing step generates the question 'How are the profiles shown in Fig. 2(a) and (b) related each to the other?'.

In response, we sought to establish a frame of reference that forms the basis for a quantitative approach. The information required is the macromolecular composition of the cells under scrutiny that can be derived from the ratios DNA : RNA : protein. The data required are not available for *M. tuberculosis* H37Rv, but are available for the very closely related family member *M. bovis* Pasteur (Beste *et al.*, 2005), which is frequently used as a model for the tubercle bacillus. These data are shown in Table S1.

The first step in our analysis was to explore the relation between $n_{\mathrm{reads(i)}}$ the number (millions) of 'reads' (partially sequenced transcripts) and $l_{tr(i)}$ nucleotides the length of the entire transcript. The subscript (i) denotes a particular ORF or operon. The plot of log $n_{\mathrm{reads(i)}}$ versus log $l_{tr(i)}$ was found to be a linear plot with a slope of 1.48, which led to the empirical result shown in equation (2). The observed numbers of reads were found to agree with the calculated values to within 10 % or better (Table 1). The plot of the number of reads (millions) versus the length of the ORF (nucleotides) raised to the power 1.48 was found to be linear (Fig. 2c).

$$n_{reads(i)} = 169 \bullet l_{ORF(i)}^{1.48} \qquad (2)$$

The slope of equation (2) is equal to the product of three components: namely, the number $n_{c-p(i)}$ of copies of the product $p_{(i)}$ of $\mathrm{ORF}_{(i)}$, the number n_{cells} of cells providing the RNA sample, and a constant κ that is characteristic of the RNaseq platform [see equation (3)].

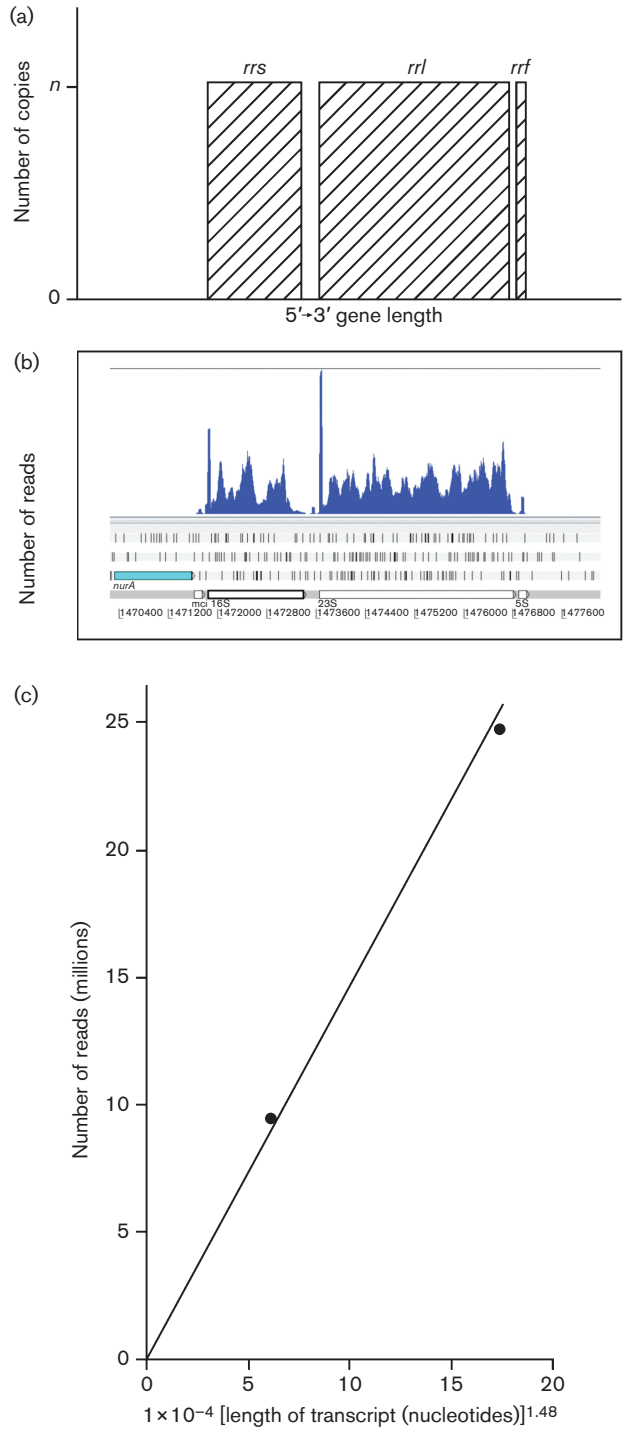

Fig. 2. Comparison of the profiles of the number of copies of *rrs*, *rrl* and *rrf* components of rRNA versus 'gene location' with the corresponding profiles of reads identified by the RNaseq method. (a) Profiles of rRNA components. (b) Profiles of the corresponding numbers of reads. (c) The relation between the numbers of reads (millions) with the length $l_{tr(i)}$ of the transcript. As discussed in the text, the plot is described by the equation $n_{reads(i)} = 169 \bullet l_{tr(i)}^{1.48}$.

$$169 = n_{c-p(i)} \bullet n_{cells} \bullet \kappa \quad (3)$$

We suppose that there are 3730 ribosomes per cell and $n_{c-p(i)}$=3730 copies for each of the three rRNA components. Hence, equation (2) may be written as equation (4), after dividing each side by 3730.

$$n_{reads(i)/p(i)} = n_{reads(i)} / 3730 = 0.045 \bullet l_{ORF(i)}^{1.48} \quad (4)$$

The term $n_{reads(i)/p(i)}$ is the number (millions) of reads of $ORF_{(i)}$ per copy of the gene product.

The rRNA species provide examples of the general case. The number $n_{tr(i)}$ of transcripts per ORF is given by the ratio $n_{reads(i)} / n_{reads(i)/p(i)}$, which is the total number of reads divided by the number of reads per copy of the product of $ORF_{(i)}$ (see equation 5).

$$n_{tr(i)} = n_{reads(i)} / n_{reads(i)/p(i)} \quad (5)$$

These equations may now be applied to the data available for both the *rpsJ* and *rplN* operons.

The number of transcripts of $ORF_{(i)}$ can be related to the number $n_{R(i)}$ of ribosomes synthesizing protein $p_{(i)}$. Transcription/translation coupling requires that each transcript is protected from degradation by nuclease action by the protective action of the bound ribosomes and their associated elongation factors when bound to mRNA plus space that is insufficient for degradosomes to bind. Provisionally, we assign one ribosome per approximately 100 nt, which is equivalent to the diameter of a ribosome (22 nm equivalent to 65 nt) flanked by 17 or so nucleotides on either side.

This conversion factor of one ribosome per 100 nt also allows us to obtain the number of transcripts per cell directly from the equation for exponential growth (see Methods); for example, see equation (1) above. Please note that equations derived in Methods are referred to as (A1) and so on. The application of equation (A5) requires the data for the macromolecular properties of cells listed in Table S1. The values obtained using equation (A5) for the numbers of transcripts per cell provide an independent test of the accuracy of the RNaseq data.

Operons *rpsJ* and *rplN*

M. tuberculosis possesses three large operons with mRNA transcripts that are comparable in length to precursor-rRNA. One operon encodes components of ATP synthase and two, the *rpsJ* and *rplN* operons, encode ribosomal proteins (Fig. S2). Properties of the 11 proteins of the *rpsJ* operon are shown in Table S2. The component proteins of the *rplN* operon are shown in Table S3. Our knowledge of *rpsJ* and *rplN* operons was gained mainly from studies of *E.*

Table 1. Comparisons of the numbers of reads observed and calculated for rRNA species

Component	Length (nt)	No. of reads (millions)		Ratio observed/calculated
		Observed	Calculated*	
rrs	1526	9.37	8.79	1.07
rrl	3137	24.67	25.18	0.98
rrf	114	0.21	0.19	1.10

*Values were calculated by using equation 2.

coli and other fast-growing bacteria. The *rpsJ* and *rplN* operons of *M. tuberculosis* and *E. coli* are homologous in structure, and we infer that they are also homologous in function.

The coding region of the *rpsJ* operon extends over 5680 bp and encodes 1803 aa. These data allow an estimate of the number of *rpsJ* transcripts per cell to be calculated by means of equation (A6). The macromolecular composition of the cells (see Table S1) allows the number of ribosomes actively translating the *rpsJ* operon at any instant to be evaluated. The same considerations apply to the *rplN* operon, which extends over 4500 bp and encodes 1349 aa (see Table S3). The numbers of transcripts per cell calculated for the two operons from RNaseq data using equation (A6) and the numbers of ribosomes estimated to be translating each of the operons using equation (A5) are presented in Table 2. The two sets of data agree to within 25 % or better. The data derived from the use of the growth equation are comparative. These data provide the basis for the schematic views of transcription (Fig. 3a) and coupled transcription/translation (Fig. 3b) of the *rpsJ* and *rplN* operons.

Protein synthesis may be viewed from several different perspectives, for example: (i) the rate $\varepsilon_{aa(i)}$ amino acids h^{-1} of peptide chain elongation; (ii) the time taken to synthesize protein $P_{(i)}$ ($\varepsilon_{aa(i)}$); and (iii) the rate (see Methods) at which completed copies of $P_{(i)}$ are produced (the 'run off' rate). It is estimated (based on $\varepsilon_{aa(i)}$ =3900 amino acids h^{-1}) that a ribosome takes 28 min to translate the *rpsJ* operon and that one operon is completed every 32 s. Similarly, 30 min are needed for a ribosome to translate the *rplN* operon and one operon is 'run off' every 32s. These properties are evident from the schematic views of transcription/translation of these operons shown in Fig. 3. Allowing for the slower growth rate of the mycobacterium, Fig. 3(a, b) are consistent with the electron micrographs (Fig. 1) of active operons visualized by Miller *et al.* (1970). The electron micrograph obtained for *E. coli* shows eight nascent transcripts compared with the estimate of approximately a single transcript for the tubercle bacillus. The numbers of ribosomes per *rplN* operon was calculated by means of equation (4). The values obtained were 45 ribosomes for *M. bovis* BCG (Fig. 3b) and 218 for *E. coli*, which is in accord with the historic electron micrograph shown in Fig. 1.

Transcriptional control elements of the *rpsJ* and *rplN* operons

The *rpsJ*, *rplN* and *rrn* operons of selected mycobacteria share features of transcriptional control that include −35 and −10 promoter elements, transcription start sites and stringent elements (Table 3), which leads us to infer that the transcriptional control elements of the three operons

Table 2. Comparison of data based on RNaseq with data derived from macromolecular composition

Property	*rpsJ* operon	*rplN* operon
Length (bp)	5680	4502
No. of amino acids	1803	1349
No. of reads (observed)	19754	13941
No. of reads (calculated)	16204	11500
No. of transcripts per cell	1.22	1.21
No. of 'bound ribosomes' [RNaseq]*	69	54
No. of 'bound ribosomes' [macromolecular composition]†	52	39

*The numbers of 'active ribosomes' are inferred values derived from RNaseq data using the approximation that there is one ribosome per 100 nt of mRNA (see text).
†Denotes comparative values calculated from data for the macromolecular composition of cells (see Table S1) by using equation (A5).

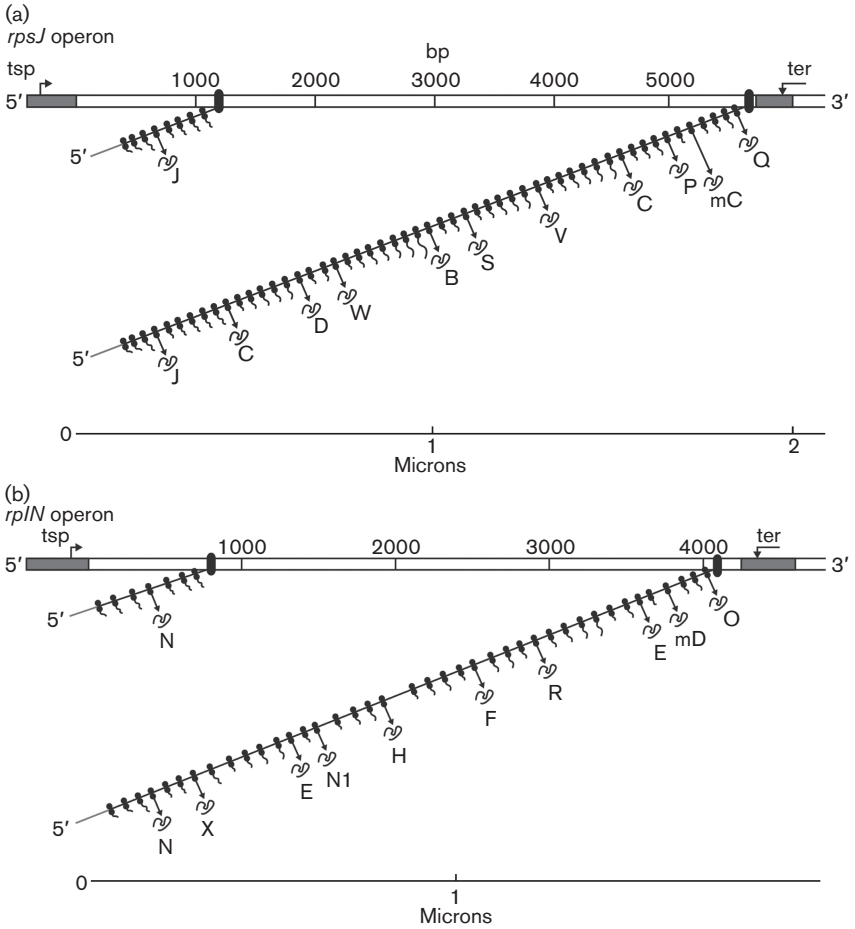

Fig. 3. Schematic views of transcription/translation of the *rpsJ* and *rplN* operons. RNAP is represented by filled cylinders with hemispherical ends and ribosomes by contiguous filled circles. Arrows show the release of completed polypeptides, each of which is represented by the same irregular shape and each is identified by a single letter. (a) *rpsJ* operon. The proteins represented are as follows: J, RpsJ; C, RplC: D, RplD; W, RplW; B, RplB; S, RpsS; V, RplV; C, RpsC; P, RplP; mC, RpmC; Q, RpsQ. (b) *rplN* operon. The proteins represented are as follows: N, RplN; X, RplX; E, RplE; N1, RpsN1; H, RpsH; F, RplF; R, RplR; E, RpsE; mD, RpmD; O, RplO.

are likely to be of comparable strengths. The ribosome binding motifs of transcripts of the *rpsJ* and *rplN* operons were found to differ (Table 3). Two strong motifs were found to be present for transcripts of *rpsJ* and only one was found in transcripts of *rplN*.

The wasteful production of an excess of the proteins encoded by the *rpsJ* and *rplN* operons is prevented by feedback control mechanisms. Inspection revealed no insight into how the *rpsJ* operon is controlled. However, a possible mechanism for the control of the *rplN* operon was discerned (Table 3). The key observation is that the motif 5′-gcagac-3′ present near the 5′ ends of transcripts of the operon match the identical sequence in mycobacterial 16S rRNA which, by analogy with *E. coli* 16S rRNA, is believed to form part of the binding site for RpsH in the 30S ribosomal subunit. Hence, we propose that when RpsH is in excess it has the capacity to bind to the initiating RNAP complex to prevent transcription of the *rplN* operon. Our scrutiny of a wide range of homologous mycobacterial sequences led us to propose a switch mechanism, which is outlined in Fig. 4. Direct experimental support for our proposal is needed.

Inspection of the genomic nucleotide sequence showed that the *rplN* operon of *M. abscessus* ATCC 19977 was found to be anomalous because the genes encoding ribosomal proteins were interrupted by five genes encoding non-ribosomal proteins. The inserted ORFs are found between *rpsN1* and *rpsH*. It is not clear how this location affects the role of RpsH in the control of the expression of the operon. The effects of the inserted genes need further study to establish the mechanisms involved in the transcription of the ribosomal protein genes and to ascertain the extent to

Table 3. Motifs of promoters and Shine–Dalgarno sequences of the operons studied

Motifs of promoters of operons encoding *rpsJ*	
M. tuberculosis H37Rv	5′-gat**ttggct**gaaagcggtgatctggggca**tactgt**tcag**g**ttgccctgcgcc...-3′
M. marinum M	5′-gat**ttggcc**gaa-ccgctgatctgaggca**tactgt**tcag**g**ttgtcttgcgcc...-3′
M. leprae Br4923	5′-aat**ttggcc**ga—-tacagcaagtagggca**tactgt**tcag**g**ttgccttgcgcc...-3′
M. smegmatis MC2 155	5′-gat**ttggag**-cacgcgcggatctagggca**tactgt**tcggg**g**ttgccttgcgcc...-3′
M. abscessus ATCC 19977	5′-gat**ttggcc**g——gcgctggataaggca**tactgt**tcggg**g**ttgccgtgaacc...-3′
Motifs of promoters of operons encoding *rplN**	
M. tuberculosis H37Rv	5′-att**ttgcgc**-tg-agtaggtcgtcccc**taaact**tcaggg**g**ttgccgtga***gcagac***ctc-***3′***
M. marinum M	5′-aat**ttggct**ctg-agcagctcgtcccc**tagact**ctaggg**g**ttgccttgg***gcagacctc-3′***
M. leprae Br4923	5′-aat**ttggct**ccg-agttggtcatcccc**taaact**ctagt**g**g**t**tgcctcgt***g**c**t**g**g**c**c**g-3′
M. smegmatis MC2 155	5′-gat**ttggag**cat-tgggcgagctcacc**tagtat**gtaggg**g**ttgccttgg***gcagac***ctt-3′
M. abscessus ATCC 19977	5′-gat**ttggg**tttcacctgcgtgttcgcc**tacact**gggccg**g**ttgtctgcg***g**t**g**g**cc***ag-3′
Motifs of promoters of *rrn* operons†	
M. tuberculosis H37Rv	5′-gtc**ttgact**ccattgccggatttgtatt**agact**ggcaggg**g**ttgccccgaagc(n)$_{27}$#-3′
M. marinum M	5′-agt**ttgact**ttcctgccggatctgtatt**aagct**ggcaggg**g**ttgccccgaaga(n)$_{29}$#-3′
M. leprae Br4923	5-′gac**ttgact**cctctgctggatctgtatt**aatct**ggctggg**g**ttccccgaagc(n)$_{27}$#-3′
M. smegmatis MC2 155	5′-gac**ttgaca**agccagacaaagcagtat**taagct**ggcaggg**g**ttgccccaaaac(n)$_{29}$#-3′
M. abscessus ATCC 19977	5′-aat**ttgact**-caggttcacgaacttga**tacggt**ttccga**g**ccgcccgaaagg(n)$_{65}$#-3′
Motif 'GCAGAC' is also located near the translation start site (underlined) of mycobacterial *rpsH* genes	
M. tuberculosis H37Rv	5′-atgacgatgacggacccgatc***gcagac***tttttgact-3′
M. marinum M	5′-atgaccatgacggacccgatc***gcagac***ttcttgact-3′
M. leprae Br4923	5′-atgaccatgacggacccgatc***gcagat***tttttgaca-3′
M. smegmatis MC2 155	5′-atgacgatgactgacccgatc***gcagac***ttcttgaca-3′
M. abscessus ATCC 1997	5′-atgaccgat————ccgatc***gcagac***tttctgaca-3′
Shine–Dalgarno motifs found in mycobacterial *rrs* operons‡	
Mycobacterial *rrs* 3′ terminal sequence	3′-ucuuuccucc-5′
Notional Shine–Dalgarno sequence	5′-agaaaggagg-3′
Observed motif	
(a)	5′-.....**gagg**-3′
(b)	5′-....**aggag**.-3′
(c)	5′-....**agga**..-3′
Shine–Dalgarno motifs found in mycobacterial *rpsJ* operons‡	
M. tuberculosis H37Rv	5′-u**gagg**ua**ggag**aagc**GUG**...-3′
M. marinum M	5′-u**gagg**ua**ggag**aagc**GUG**...-3′
M. leprae Br4923	5′-u**gagg**ua**gga**aaagc**GUG**...-3′
M. smegmatis MC2 155	5′-c**cagg**ua**ggag**aagc**GUG**...-3′
M. abscessus ATCC 19977	5′-gua**gagg**ua**gga**agc**GUG**...-3′
Shine–Dalgarno motifs found in mycobacterial *rplN* operons‡	
M. tuberculosis H37Rv	5′-..aggtc**aggag**aucua**GUG**...-3′
M. marinum M	5′-..gaguc**aggag**aucua**GUG...-3′**
M. leprae Br4923	5′-**agga**aucgcgacugaag**GUG...-3′**
M. smegmatis MC2 155	5′-..agguc**aggag**aucga**GUG...-3′**
M. abscessus ATCC 19977	5′-..acgu**aggag**agacca**GUG...-3′**

*The motif shown in italics is a potential binding site for RpsH.
†The hash sign denotes the 12 base motif (5′-gagaactcaata-3′) that forms part of the RNase III binding site.
‡The observed motifs are shown in bold and are underlined. The termination codon is shown in bold upper-case letters.

which the rates of synthesis of the ribosomal proteins are affected.

Role of NusG in transcription/translation coupling

Nus factors (N utilizing substances) NusA, NusB, NusE (RpsJ) and NusG were first identified in *E. coli* infected with bacteriophage lambda. These host factors were found to play essential roles in the expression of protein N of the bacteriophage and they are also required for the expression of each of the three operons under scrutiny (Burmann *et al.*, 2010). For example, NusG is a factor that is essential for transcription/translation coupling. This factor has three separate domains and the functions of two of them are known. The NusG N-terminal domain (NusG-NTD) has the capacity to bind to RNAP, whereas the C-terminal domain (NusG-CTD) can combine with the NusE (RpsJ) component of ribosomes. These two functions of NusG

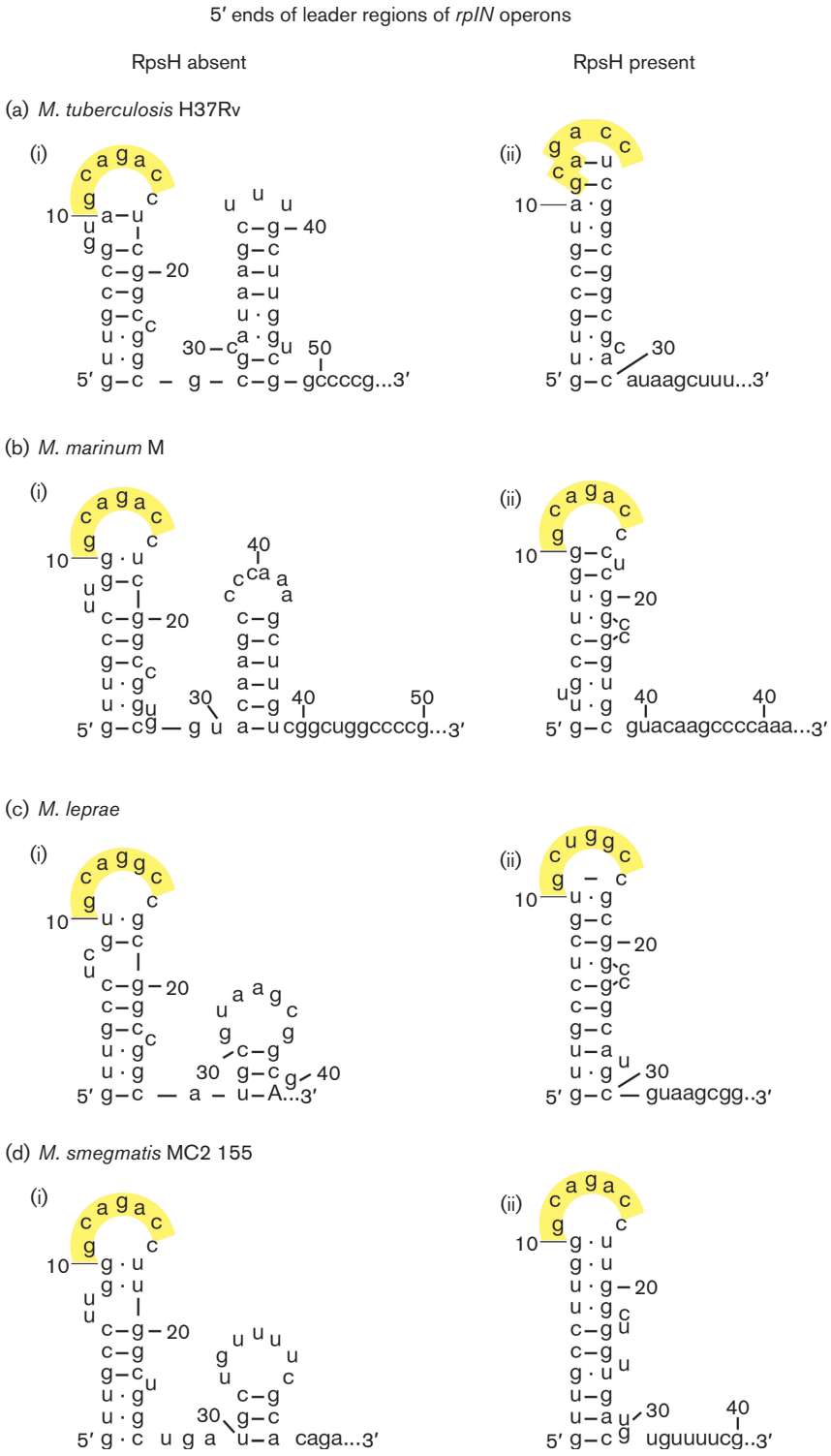

Fig. 4. Possible effects of RpsH binding on transcription of *rplN* operons. The motif 5′-gcagac-3′ (highlighted) forms part of the 16S rRNA (*rrs*) binding site for RpsH. Hence, we infer that the presence of this motif near to the 5′ end of transcripts of the *rplN* operon is also part of a potential binding site for RpsH. We propose that transcription proceeds freely when concentrations of unbound RpsH are low (i) and that increasing concentrations of free RpsH may lead to recruitment of RpsH by the RNAP complex leading to termination of transcription (ii). The proposed switch is common to *Mycobacterium* since similar structures were found to be present at the 5′ ends of the *rplN* operons of different mycobacteria (see Table 3).

enable transcription to be coupled with translation. NusG-CTD can also bind to Rho to terminate transcription (Burmann *et al.*, 2010; McGary & Nudler, 2013; Proshkin *et al.*, 2010). The rate of transcription depends on the rate of translation because the ribosome moves through a progressive, unilateral, translocation (Fig. 5). In contrast, in the absence of NusG, RNAP is capable of backtracking within the transcription bubble. When transcription and translation are coupled, the movement of the ribosome modulates the rate of transcription by preventing backtracking and thereby increasing the efficiency of transcription (Proshkin *et al.*, 2010).

Two factors, NusB and NusE (RpsJ) carry out an essential role in the transcription of *rrn* operons by preventing premature termination of the transcript. The crystal structure of NusB from *M. tuberculosis* is known (Gopal *et al.*, 2000a) and its interaction of with NusE (RpsJ) has been investigated (Gopal *et al.*, 2000b)

Concluding remarks

The electron micrographs of Miller *et al.* (1970) showed the presence of RNA transcripts longer than a micron transcribed from a section of DNA corresponding in size to the *rplN* operon, and also showed that transcription and translation were coupled. As described above, NusG is the factor required for this coupling to take place. We have presented a quantitative view of factors affecting the synthesis of two operons, *rplN* and *rpsJ*, encoding a total of 21 ribosomal proteins. Quantitative analysis was based on two approaches, namely: the mathematical framework developed from the equation for exponential growth and RNaseq measurements. Ribosomal proteins RplD, RpsH and RpsJ were found to participate in regulating the expressions of the *rpsJ*, *rplN* and the rRNA operons, respectively. The unusual structure of the *rplN* operon of *M. abscessus* ATCC 19977 requires further investigation. Transcription start sites of the *rpsJ* and *rplN* operons of *M. tuberculosis* H37Rv were first identified by an application of the RNaseq procedures (Cortes *et al.*, 2013) and homologous sequences were then found in four other representative mycobacterial species, namely: the pathogens *M. marinum* and *M. leprae* Br4923, and the opportunistic pathogens *M. smegmatis* MC2 155 (Brown-Elliott & Wallace, 2002) and *M. abscessus* ATCC 19977. The comparison aids the identification of functional motifs based on the notion that sequences diverge during evolution unless they are constrained by functional requirements. The results contribute to our understanding of the early stages of ribosome synthesis during mycobacterial growth.

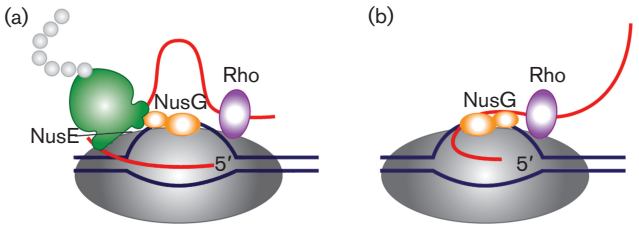

Fig. 5. The roles of NusG in transcription/translation coupling. (a) Composition of an active RNAP complex. RNAP is shown in dark grey, DNA in blue and nascent RNA in red. The ribosome is shown in green with the nascent polypeptide chain in light grey; the bulge in the small subunit denotes the location of NusE (RpsJ). NusG is shown in orange: its shape denotes two functional sections. The larger section denotes the N-terminal domain, which binds to RNAP. The smaller section denotes the C-terminal domain, which interacts with NusE *in situ*. Rho is shown in purple. (b) After translation is completed NusG remains bound to RNAP and may also bind to Rho through the C-terminal domain leading to termination of transcription.

ACKNOWLEDGEMENTS

We thank Professor Douglas Young for his interest and encouragement, and our colleague Dr Willie Taylor for helpful discussions. We thank vertis Biotechnologie (Germany) for cDNA library preparation and sequencing. This work was supported by European Community grant SysteMTb HEALTH-F4-2010-241587 FP7 and UK Medical Research Council grant U117581288.

REFERENCES

Arnvig, K. B., Comas, I., Thomson, N. R., Houghton, J., Boshoff, H. I., Croucher, N. J., Rose, G., Perkins, T. T., Parkhill, J. & other authors (2011). Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. *PLoS Pathog* **7**, e1002342.

Bashan, A. & Yonath, A. (2008). Correlating ribosome function with high-resolution structures. *Trends Microbiol* **16**, 326–335.

Beste, D. J., Peters, J., Hooper, T., Avignone-Rossa, C., Bushell, M. E. & McFadden, J. (2005). Compiling a molecular inventory for *Mycobacterium bovis* BCG at two growth rates: evidence for growth rate-mediated regulation of ribosome biosynthesis and lipid metabolism. *J Bacteriol* **187**, 1677–1684.

Brown-Elliott, B. A. & Wallace, R. J., Jr (2002). Clinical and taxonomic status of pathogenic nonpigmented or late-pigmenting rapidly growing mycobacteria. *Clin Microbiol Rev* **15**, 716–746.

Burmann, B. M., Luo, X., Rösch, P., Wahl, M. C. & Gottesman, M. E. (2010). Fine tuning of the *E. coli* NusB : NusE complex affinity to BoxA RNA is required for processive antitermination. *Nucleic Acids Res* **38**, 314–326.

Byrne, R., Levin, J. G., Bladen, H. A. & Nirenberg, M. W. (1964). The in vitro formation of a DNA-ribosome complex. *Proc Natl Acad Sci U S A* **52**, 140–148.

Coenye, T. & Vandamme, P. (2005). Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. *FEMS Microbiol Lett* **242**, 117–126.

Cortes, T., Schubert, O. T., Rose, G., Arnvig, K. B., Comas, I., Aebersold, R. & Young, D. B. (2013). Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Rep* **5**, 1121–1131.

Cox, R. A. (2004). Quantitative relationships for specific growth rates and macromolecular compositions of *Mycobacterium tuberculosis*,

Streptomyces coelicolor A3(2) and *Escherichia coli* B/r: an integrative theoretical approach. *Microbiology* **150**, 1413–1426.

Cox, M. P., Peterson, D. A. & Biggs, P. J. (2010). SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**, 485.

Gonzalez-y-Merchand, J. A., Colston, M. J. & Cox, R. A. (1996). The rRNA operons of *Mycobacterium smegmatis* and *Mycobacterium tuberculosis:* comparison of promoter elements and of neighbouring upstream genes. *Microbiology* **142**, 667–674.

Gonzalez-y-Merchand, J. A., Garcia, M. J., Gonzalez-Rico, S., Colston, M. J. & Cox, R. A. (1997). Strategies used by pathogenic and nonpathogenic mycobacteria to synthesize rRNA. *J Bacteriol* **179**, 6949–6958.

Gopal, B., Cox, R. A., Colston, M. J., Dodson, G. G., Smerdon, S. J. & Haire, L. F. (2000a). Crystallization and preliminary X-ray diffraction studies on the N-utilizing substance-B (NusB) from *Mycobacterium tuberculosis. Acta Crystallogr D Biol Crystallogr* **56**, 64–66.

Gopal, B., Haire, L. F., Cox, R. A., Colston, M. J., Major, S., Brannigan, J. A., Smerdon, S. J. & Dodson, G. (2000b). The crystal structure of NusB from *Mycobacterium tuberculosis. Nat Struct Biol* **7**, 475–478.

Ji, Y., Colston, M. J. & Cox, R. A. (1994a). The ribosomal RNA (*rrn*) operons of fast-growing mycobacteria: primary and secondary structures and their relation to *rrn* operons of pathogenic slow-growers. *Microbiology* **140**, 2829–2840.

Ji, Y., Colston, M. J. & Cox, R. A. (1994b). Nucleotide sequence and secondary structures of precursor 16S rRNA of slow-growing mycobacteria. *Microbiology* **140**, 123–132.

Ji, Y., Kempsell, K. E., Colston, M. J. & Cox, R. A. (1994c). Nucleotide sequences of the spacer-1, spacer-2 and trailer regions of the *rrn* operons and secondary structures of precursor 23S rRNAs and precursor 5S rRNAs of slow-growing mycobacteria. *Microbiology* **140**, 1763–1773.

Kempsell, K. E., Ji, Y., Estrada-G, I. C. E., Colston, M. J. & Cox, R. A. (1992). The nucleotide sequence of the promoter, 16S rRNA and spacer region of the ribosomal RNA operon of *Mycobacterium tuberculosis* and comparison with *Mycobacterium leprae* precursor rRNA. *J Gen Microbiol* **138**, 1717–1727.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760.

Mattheakis, L., Vu, L., Sor, F. & Nomura, M. (1989). Retroregulation of the synthesis of ribosomal proteins L14 and L24 by feedback repressor S8 in *Escherichia coli. Proc Natl Acad Sci U S A* **86**, 448–452.

McGary, K. & Nudler, E. (2013). RNA polymerase and the ribosome: the close relationship. *Curr Opin Microbiol* **16**, 112–117.

Miller, O. L., Jr & Hamkalo, B. A. (1972). *Functional Units in Protein Biosynthesis* (FEBS Symposium no. 23). London: Academic Press.

Miller, O. L., Jr, Hamkalo, B. A. & Thomas, C. A., Jr (1970). Visualization of bacterial genes in action. *Science* **169**, 392–395.

Proshkin, S., Rahmouni, A. R., Mironov, A. & Nudler, E. (2010). Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **328**, 504–508.

Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.

Ripoll, F., Pasek, S., Schenowitz, C., Dossat, C., Barbe, V., Rottman, M., Macheras, E., Heym, B., Herrmann, J. L. & other authors (2009). Non mycobacterial virulence genes in the genome of the emerging pathogen *Mycobacterium abscessus. PLoS ONE* **4**, e5660.

Schaechter, M., Maaloe, O. & Kjeldgaard, N. O. (1958). Dependency on medium and temperature of cell size and chemical composition during balanced grown of *Salmonella typhimurium. J Gen Microbiol* **19**, 592–606.

Stelzl, U., Zengel, J. M., Tovbina, M., Walker, M., Nierhaus, K. H., Lindahl, L. & Patel, D. J. (2003). RNA-structural mimicry in *Escherichia coli* ribosomal protein L4-dependent regulation of the S10 operon. *J Biol Chem* **278**, 28237–28245.

Stent, G. S. (1964). The operon: on its third anniversary. Modulation of transfer RNA species can provide a workable model of an operator-less operon. *Science* **144**, 816–820.

Tobin, D. M. & Ramakrishnan, L. (2008). Comparative pathogenesis of *Mycobacterium marinum* and *Mycobacterium tuberculosis. Cell Microbiol* **10**, 1027–1039.

Williamson, J. R. (2009). The ribosome at atomic resolution. *Cell* **139**, 1041–1043.

Edited by: S. Gordon