DNA Research 18, 413–422, (2011)
Advance Access Publication on 14 November 2011

doi:10.1093/dnares/dsr028

# *Mycoplasma hyopneumoniae* Transcription Unit Organization: Genome Survey and Prediction

Franciele Maboni Siqueira, Augusto Schrank, and Irene Silveira Schrank*

*Programa de Pós-Graduação em Biologia Molecular e Celular, Departamento de Biologia Molecular e Biotecnologia, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil*

*To whom correspondence should be addressed. Tel. +55 51-3308-6055. Fax. +55 51-3308-7309. Email: irene@cbiot.ufrgs.br

Edited by Naotake Ogasawara
(Received 23 March 2011; accepted 24 July 2011)

## Abstract

***Mycoplasma hyopneumoniae* is associated with swine respiratory diseases. Although gene organization and regulation are well known in many prokaryotic organisms, knowledge on mycoplasma is limited. This study performed a comparative analysis of three strains of *M. hyopneumoniae* (7448, J and 232), with a focus on genome organization and gene comparison for open read frame (ORF) cluster (OC) identification. An *in silico* analysis of gene organization demonstrated 117 OCs and 34 single ORFs in *M. hyopneumoniae* 7448 and J, while 116 OCs and 36 single ORFs were identified in *M. hyopneumoniae* 232. Genomic comparison revealed high synteny and conservation of gene order between the OCs defined for 7448 and J strains as well as for 7448 and 232 strains. Twenty-one OCs were chosen and experimentally confirmed by reverse transcription–PCR from *M. hyopneumoniae* 7448 genome, validating our prediction. A subset of the ORFs within an OC could be independently transcribed due to the presence of internal promoters. Our results suggest that transcription occurs in 'run-on' from an upstream promoter in *M. hyopneumoniae*, thus forming large ORF clusters (from 2 to 29 ORFs in the same orientation) and indicating a complex transcriptional organization.**

**Key words:** ORF cluster; intergenic regions; cotranscription; transcriptional units

## 1. Introduction

*Mycoplasma hyopneumoniae* is the key aetiological agent of porcine enzootic pneumonia (EP) and is a major contributor to porcine respiratory disease complex.[1] EP is a chronic respiratory disease that mainly affects finishing pigs. Although major efforts to control *M. hyopneumoniae* infection and its detrimental effects have been made, swine production has suffered from significant economic loss.

Like other mycoplasmas, *M. hyopneumoniae* has a small genome with limited biosynthetic potential.[2–4] Up until now, four *M. hyopneumoniae* genome strains have been sequenced.[2–4] Despite the significant amount of data produced by genome sequencing, information on its open read frame (ORF) cluster (OC) organization, transcriptional unit (TU) formation and transcriptional regulation is very limited. Genes organized in an ORF cluster (OC) or TU are arranged in tandem in the genomic sequence, being delimited by the location of the upstream promoter and the downstream transcriptional terminator. TU identification is commonly based on bioinformatics approaches combined with experimental data. Mycoplasma studies have demonstrated the occurrence of polycistronic mRNA and TUs in some genomic regions.[5–12] However, information on characterized TUs is available for exceptional cases, making it difficult to judge the accuracy of current *in silico* TU or OC prediction methods in mycoplasma genome sequences.

Prediction and recognition of mycoplasma promoter elements are poorly developed. The unusual A + T content of intergenic regions (IRs) in mycoplasma[13] and the weak −35 consensus[14] prevent efficient prediction of promoters by current bioinformatics approaches. Moreover, in spite of the development of artificial transformation and other genetic tools for some species of *Mycoplasma*,[14–17] these methodologies are yet to be developed for *M. hyopneumoniae*. Therefore, strategies for the mapping of genome organization and gene transcription in *M. hyopneumoniae* can help understand regulatory mechanisms and form the basis for future work on mapping promoters and terminators.

Information regarding transcription promoter or terminator sequences have yet to be defined in *M. hyopneumoniae*. Therefore, any analyses at the level of transcription regulation are difficult. Recently, Gardner and Minion[18] demonstrated that transcription occurs in IRs of *M. hyopneumoniae*, indicating that *M. hyopneumoniae* does not strictly control transcription termination. An analysis of the average RNA folding energy near stop codons suggested that no stem-loops are formed in *Mycoplasma genitalium* and *Mycoplasma pneumoniae* TUs, indicating the existence of qualitatively different and uncharacterized mechanisms for transcription termination.[19] Defining a transcription termination region in mycoplasmas has revealed some contradictory results. In some operons of *M. genitalium*,[8] the presence of stem-loop structures have been found, and results from Hoon *et al*.[20] suggest that the Rho-independent terminator represents the main mode of transcriptional termination in mycoplasmas. However, it has been suggested that genes are arranged in long clusters in mycoplasmas, indicating the possibility that the starting point of transcription occurs in the first gene and stops at low frequencies due to the absence of specific regions for transcription termination.[21]

Although gene co-regulation occurs in mycoplasmas, to date only little information is available on ORF organization in the *M. hyopneumoniae* genome, as well as its relationship with pathogenicity. In order to establish differences between the gene organization of pathogenic and non-pathogenic *M. hyopneumoniae* and also to relate this gene organization to transcription regulatory mechanisms, we systematically analysed the synteny and conservation of ORF order among three *M. hyopneumoniae* genomes (two pathogenic strains: 7448 and 232; one non-pathogenic strain: J). We then performed the experimental characterization of some OCs predicted in the *M. hyopneumoniae* 7448 strain. This is the first report where gene organization analysis was systematically performed and compared in three *M. hyopneumoniae* genomes, two of which are pathogenic to swines and a significant number of predicted OCs are experimentally demonstrated and named TUs. These results could contribute to promote the understanding of *M. hyopneumoniae* transcription regulation.

## 2. Materials and methods

### 2.1. In silico *analysis of TUs*

OC prediction was performed by Artemis Release 10.5.2 software.[22] The criteria for the manual examination of possible OCs in *M. hyopneumoniae* genomes (MHP7448_GenBank: NC_007332; MHP232_GenBank: NC_006360 and MHPJ_GenBank: NC_007295) were established on the occurrence of clusters with two or more genes in tandem in the same DNA strand. This was performed by systematic annotation comparison of the protein sequences encoded in all ORFs from the analysed genomes. According to the complexity of adjacent ORF rearrangements, two groups were created: OCs characterized by the presence of two or more ORFs in the same DNA strand until the occurrence of ORFs in the opposite strand, and monocistronic (mC) group representing single ORFs. Differences in the annotation among the genomes were evaluated by comparing protein sequences using the NCBI/BLASTP program.

The comparative analysis of OC organization found among the three *M. hyopneumoniae* genomes (7448, 232 and J) was accomplished based on the *M. hyopneumoniae* 7448 genome results by Artemis Comparison tools Release 9.0.[23]

### 2.2. *Primer design*

The primers were designed in Vector NTI Advance 10 (Invitrogen) based on the *M. hyopneumoniae* 7448 genome sequence (GenBank: NC_007332). Primer pairs were chosen to enable synthesis by reverse transcriptase–PCR (RT–PCR), whose products span each IR, demonstrating the occurrence of polycistronic mRNA and confirming the prediction previously made. The $OC_{7448}111$ (Fig. 1) was used as an example where three primer pairs were designed for utilization in the RT–PCR analysis. As shown in Fig. 1, primer pairs were distributed in the predicted OC and used in the following combination in the RT–PCRs: 1°F/1°R; 2°F/2°R; 3°F/3°R; 1°F/2°R and 2°F/3°R. To generate experimental data, a total of 71 primer pairs were projected and employed with this methodology (Supplementary Table S1).

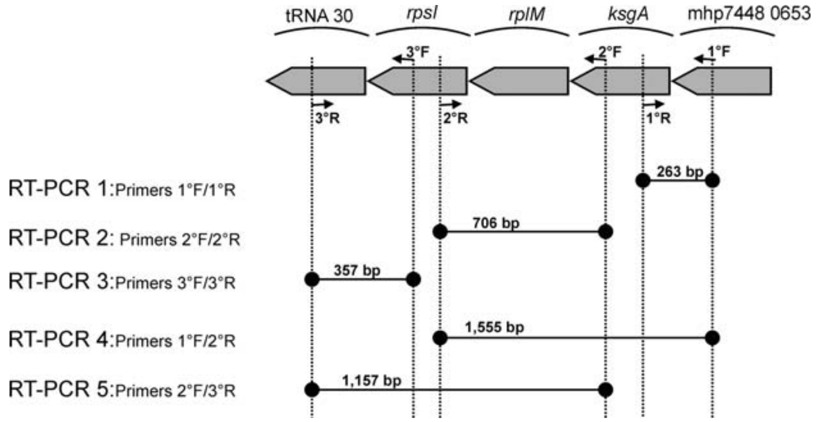

**Figure 1.** OCs in *M. hyopneumoniae*. A series of RT–PCRs using ORF-specific primers are designed to produce the products shown at the centre of the figure (black lines with the product size in bp). The black arrows 1°F/1°R, 2°F/2°R and 3°F/3°R correspond to primers 111 1°F, 111 1°R, 111 2°F, 111 2°R, 111 3°F and 111 3°R, respectively, and represent a schematic localization of primers pairs in OC$_{7448}$111 designed to span the IRs between ORFs. The ORFs and tRNA30 nomenclature is in agreement with the GenBank NC_007332 and are represent by grey arrows.

### 2.3. *DNA and RNA manipulations*

*Mycoplasma hyopneumoniae* strain 7448 was grown in 5 and 25 ml of Friis broth[24] at 37°C for 48 h for DNA extraction and RNA isolation, respectively. Genomic DNA of *M. hyopneumoniae* was isolated using standard protocols.[25] The quality of the isolated DNA was verified by gel electrophoresis and displayed pure high-molecular-weight DNA.

Total RNA of *M. hyopneumoniae* extraction was performed with Trizol® (Invitrogen) following the manufacturer's guidelines, including 1 U/μg DNase/RNAse free (Promega) treatment, and then the extracted RNA was quantified in Qubit™ system (Invitrogen) and also analysed by gel electrophoresis. The RNA extraction was stored at −70°C.

### 2.4. *Reverse transcriptase–PCR*

A first-strand cDNA synthesis reaction was conducted by adding 1 μg of total RNA to 132.5 ng of pd(N)$_6$ random hexamer (GE Healthcare) and 10 mM deoxynucleotide triphosphates. The mixture was heated to 65°C for 5 min and then incubated on ice for 1 min. First-strand buffer (250 mM Tris–HCl, pH 8.3, 375 mM KCl, 15 mM $MgCl_2$), 0.1 M dithiothreitol, 40 U RNase inhibitor (Invitrogen) and 200 U M-MLV RT (Moloney Murine Leukemia Virus Reverse Transcriptase—Invitrogen) was then added to a total volume of 20 μl. The reaction mixture was incubated at 25°C for 10 min and at 37°C for 50 min followed by 15 min at 70°C. Negative control was prepared in parallel, differing only by the absence of the RT enzyme.

PCRs included 1 U *Taq* DNA polymerase (Cenbiot Enzimas), 10× buffer (100 mM Tris–HCl, pH 8.5, 500 mM KCl), 1 mM magnesium chloride, 100 mM deoxynucleotide triphosphates, 10 pmol of each primer (Supplementary Table S1) and 1 μl of the first-strand cDNA reaction in a final volume of 25 μl. Negative control of RT–PCR was prepared in parallel, which differed only by the absence of cDNA, and no genomic DNA was added to the reaction mixture for the PCR control. PCR positive control was prepared using the genomic DNA of *M. hyopneumoniae* as a template. The PCR conditions were as follows: 1 cycle at 94°C for 5 min was followed by 30 cycles of 94°C for 30 s; denaturation and extension temperature and time varied according to each primer pair (Supplementary Table S1). The final extension step was at 72°C for 10 min. Reaction products were analysed in 1.2% agarose gels and fragments of the expected size were precipitated with tRNA (Invitrogen) as follows. To each RT–PCR (22 μl), 1 μg of tRNA and 2.5 V of absolute ethanol (4°C) were added and then incubated for 16 h at −20°C. After 30 min of centrifugation and wash with 70% ethanol, the pellets were resuspended in 15 μl of ultrapure water and sequenced with the DYEnamic ET Dye Terminator Cycle Sequencing Kit (Amersham Biosciences) in the MegaBACE 1000, as recommended by the manufacturer.

## 3. Results and discussion

### 3.1. *Genes in* M. hyopneumoniae *are preferably organized in OCs*

This study represents a global assessment of OC organization with *in silico* and *in vitro* analysis of the *M. hyopneumoniae* genome. Our detailed analysis indicated that genes are organized preferably in OCs. The gene-by-gene genome organization of the three strains of *M. hyopneumoniae* was analysed, gene localization was compared to detect ORFs with order conservation and two groups were created as described in Materials and methods: the OC group (named OC$_{7448}$1–OC$_{7448}$117 in the 7448 strain; OC$_{J}$1–OC$_{J}$117 in the J strain and OC$_{232}$1–OC$_{232}$116 in the 232 strain) and the mC group (named mC$_{7448}$1–mC$_{7448}$34 in the 7448 strain; mC$_{J}$1–mC$_{J}$34 in the J strain and mC$_{232}$1–mC$_{232}$36 in the 232 strain). In summary, 117 OCs and 34 mCs (strains 7448 and J) and 116 OCs and 36 mCs (strain 232) were detected (Table 1). The mC ORFs of the three *M. hyopneumoniae* strains are listed in Supplementary Table S2. Surprisingly, just 5% of the ORFs in the three genomes were mC. Therefore, the majority (95%) of *M. hyopneumoniae* ORFs are transcribed in long polycistronic mRNAs, thus forming extensive OCs. Moreover, the ORF numbers that constitute each OC was very variable (Table 2). As shown in Table 2, the majority of the OCs were composed of two ORFs (22%), with increasing numbers of up to 29 ORFs. Interestingly, in the three genomes analysed, approximately 85% of the OCs were composed of two to eight ORFs, demonstrating the presence of OCs with fewer numbers of ORFs.

**Table 1.** General feature of the OCs' organization in *M.hyopneumoniae* genomes

| Feature | MHP7448 | MHP J | MHP232 |
|---|---|---|---|
| Total length (base pairs) | 920 079 | 897 405 | 892 758 |
| Total no. protein coding ORFs | 657 | 657 | 691 |
| Total no. of predicted OCs | 117 | 117 | 116 |
| Total no. of mC (% compared with total ORFs) | 34 (5.1%) | 34 (5.1%) | 36 (5.2%) |
| Valid ORFs | 17 | 12 | 17 |
| Conserved hypothetical ORFs | 8 | 7 | 10 |
| Hypothetical ORFs | 9 | 15 | 9 |
| OCs organization similarity (%)[a] | 100% | 84% | 71.00% |
| OCs organization similarity (%) without RIs[b] | 100% | 85.0% | 82.0% |
| Location of region with inverted sequence (bp) | 101 560–354 533 | Not present | 104 691–344 079 |
| Transposases + ICEH-II[c] | 9 + 1 | 7 + 0 | 5 + 1 |
| Validated predicted OCs (% compared with total predicted OCs) | 21 (18%) | 0 | 0 |

[a] Similarity evaluated in relationship with MHP7448. The percentage represents the total OCs conservation (equality ORFs and RIs).

[b] Similarity evaluated in relationship with MHP7448. The percentage represents the total OCs conservation leave out RIs differences.

[c] Probably relationeted with sites of recombination in comparative analysis of OC organization between 7448 strain and 232 or J strains.

Genome minimization can be explained by reducing the number of genes or by packing them into strings of clustered genes. In the genus Mycoplasma, during genome reduction both processes seem to have occurred. In *M. hyopneumoniae*, the genome has been reduced by a loss of genes as those related with the ability to formylate Met-tRNAi[3] among others. Moreover, as shown in Table 2, the density of genes in OC$_{7448}$7 and OC$_{7448}$102 could be explained by the genome compaction. Our results in Table 2 are in agreement with those reported by Rogozin *et al.*,[26] suggesting that, in diverse organisms, the majority of gene clusters are formed by a strings of two to four genes. Therefore, this arrangement found in the three *M. hyopneumoniae* strains could be the result of the maintenance of the size of the gene clusters during genome reduction.

### 3.2. *High level of synteny of the OCs is found among* M. hyopneumoniae *strains*

Detailed analysis of each OC revealed a high level of synteny between the OCs defined for the 7448 and J strains as well as for the 7448 and 232 strains. Gene order was conserved with 85% similarity when comparing the 7448 and J strains, and 82% when comparing the 7448 and 232 strains (Table 1, Supplementary Tables S3–S5). Previous comparative analysis of these

**Table 2.** ORF composition in OCs of *M. hyopneumoniae* genomes

| Feature | MHP7448 | MHP J | MHP232 |
|---|---|---|---|
| Total no. of OCs | 117 | 117 | 116 |
| OCs composed of two ORFs | 26 | 26 | 25 |
| OCs composed of three ORFs | 19 | 18 | 16 |
| OCs composed of four ORFs | 16 | 14 | 16 |
| OCs composed of five ORFs | 15 | 16 | 13 |
| OCs composed of six ORFs | 9 | 12 | 9 |
| OCs composed of seven ORFs | 11 | 10 | 10 |
| OCs composed of eight ORFs | 5 | 5 | 5 |
| OCs composed of nine ORFs | 2 | 1 | 4 |
| OCs composed of 10 ORFs | 3 | 4 | 6 |
| OCs composed of 11 ORFs | 3 | 4 | 2 |
| OCs composed of 12 ORFs | 2 | 2 | 2 |
| OCs composed of 13 ORFs | 1 | 1 | 3 |
| OCs composed of 15 ORFs | 1 | 1 | 2 |
| OCs composed of 20 ORFs | 1 | 0 | 0 |
| OCs composed of 21 ORFs | 0 | 0 | 1 |
| OCs composed of 22 ORFs | 1 | 0 | 0 |
| OCs composed of 24 ORFs | 0 | 1 | 0 |
| OCs composed of 26 ORFs | 1 | 1 | 0 |
| OCs composed of 27 ORFs | 0 | 0 | 1 |
| OCs composed of 29 ORFs | 1 | 1 | 1 |

three genomes was performed at a genomic scale and strain-specific differences were demonstrated, but high rates of conservation were detected in regions of rearrangements, which were probably involved in pathogenesis.[3] In agreement with Vasconcelos *et al.*,[3] our work also demonstrated that the prevalence of ORFs transcribed in the DNA strand with conserved OCs reflected the high level of synteny among the strains' genomes. This occurs not only due to the same arrangement of ORFs but, more interestingly, also due to the presence of IRs of the same size. Although most of the strains display high synteny, low-degree conservation regions also occur. The analysis of genome regions with lower synteny evidences differences in ORF organization and size of intergenic distance. Some cases with common differences are shown in Fig. 2 and are related to: (Fig. 2A) ORFs that are separated into different OCs (for example OC$_{232}$46, OC$_{7448}$10 and OC$_{7448}$47, Supplementary Tables S3 and S5); (Fig. 2B) addition or removal of OCs (see OC$_{232}$02 and mC$_{232}$01, mC$_{7448}$01 and mC$_{7448}$02, Supplementary Tables S2, S3 and S5); (Fig. 2C) OCs with addition or removal of ORFs, generally encoding hypothetic proteins (for example OC$_{7448}$06 and OC$_{232}$07 or OC$_{7448}$100 and OC$_{J}$100, Supplementary Tables S3–S5).

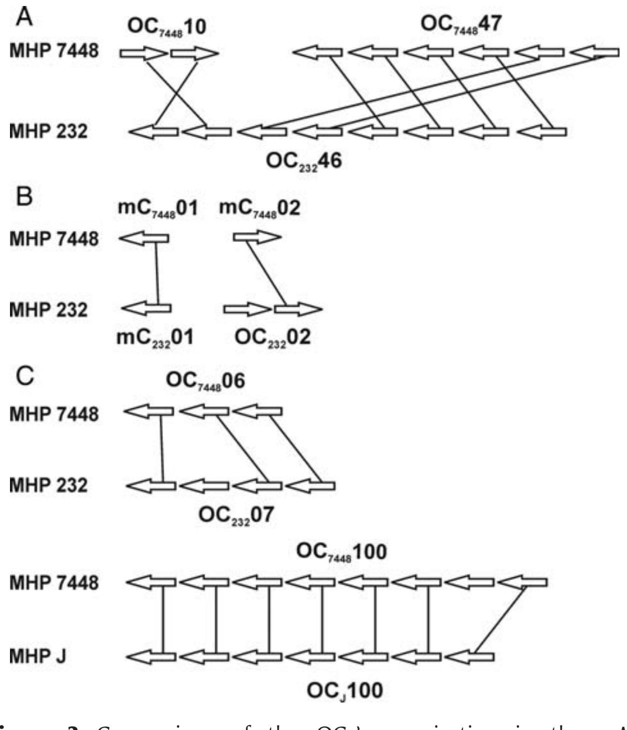

**Figure 2.** Comparison of the OCs' organization in three *M. hyopneumoniae* strains. (**A**) Homologous ORFs that are separated into different OCs in MHP7448 strain. (**B**) Addition or removal of OCs: the homologous ORF is mC in 7448 strain and is polycistronic in 232 strain. (**C**) OCs with addition or removal of ORFs, generally encoding hypothetic proteins. The bars connecting the arrows represent close matches. White arrowheads represent the ORFs and indicate the direction of transcription.

In *M. hyopneumoniae*, the overall distribution of ORFs within the OCs is highly variable in the number of ORFs and size of the intergenic distances, as well as in the functional categories of the encoded products. Approximately 92% of the OCs display some variations related to these parameters (see Supplementary Tables S3–S5). However, some of the OCs are composed of ORFs whose products are related functionally, some examples are: OC$_{7448}$22, OC$_{7448}$43, OC$_{7448}$86, OC$_{7448}$105 and OC$_{7448}$115, all OCs with orthologs in the 232 and J strains (Supplementary Tables S3–S5). Taken together, these results suggest that the classical definition of operon cannot be applied to *M. hyopneumoniae* genomes.

Despite the high level of synteny between the genomes of *M. hyopneumoniae* 7448, J and 232, strain-specific differences, such as inverted regions and rearrangements, were demonstrated, being potentially related to pathogenesis.[3] The inverted segment present in the 232 strain was further analysed and compared with the cognate 7448 region, taking into consideration the ORF string organization and OC distribution (Fig. 3). In *M. hyopneumoniae* 232, this region corresponds to genome positions from nucleotides 104 691–353 186 bp (OC$_{232}$11–OC$_{232}$46), with a total length of 248 495 bp. In *M. hyopneumoniae* 7448, the equivalent location is from nucleotides 101 560–354 533 bp (OC$_{7448}$10–OC$_{7448}$46), with a size of 252 973 bp. In this location, the overall conservation of gene order and OC distribution were preserved (Fig. 3). Typical examples of this organization are OC$_{7448}$28 and OC$_{232}$28, each containing 29 ORFs. Other examples are OC$_{232}$15 and OC$_{7448}$41 with 12 conserved ORFs each (see Supplementary Tables S3 and S5). A closer inspection of the regions bordering these inverted genome segments showed that, in each case, OC conservation was also preserved (see OC$_{232}$10 and OC$_{7448}$9). Therefore, the genome was inverted but the ORF organization was maintained. Moreover, at the right border side, the distribution of the ORF order was maintained, even when OC$_{232}$46 split from OC$_{7448}$10 and OC$_{7448}$47. It should be noted that the inversion event did not change the relative ORF order in the OCs. Taken together, these results demonstrate the similarity in ORF strings among the three strains, further supporting the hypotheses of Chen *et al.*[27] Price *et al.*[28] and Roback *et al.*[29] that the conservation in OC or TU organization in two or more phylogenetically related species is an important criterion for OC and after TU determination. Therefore, our results suggest that

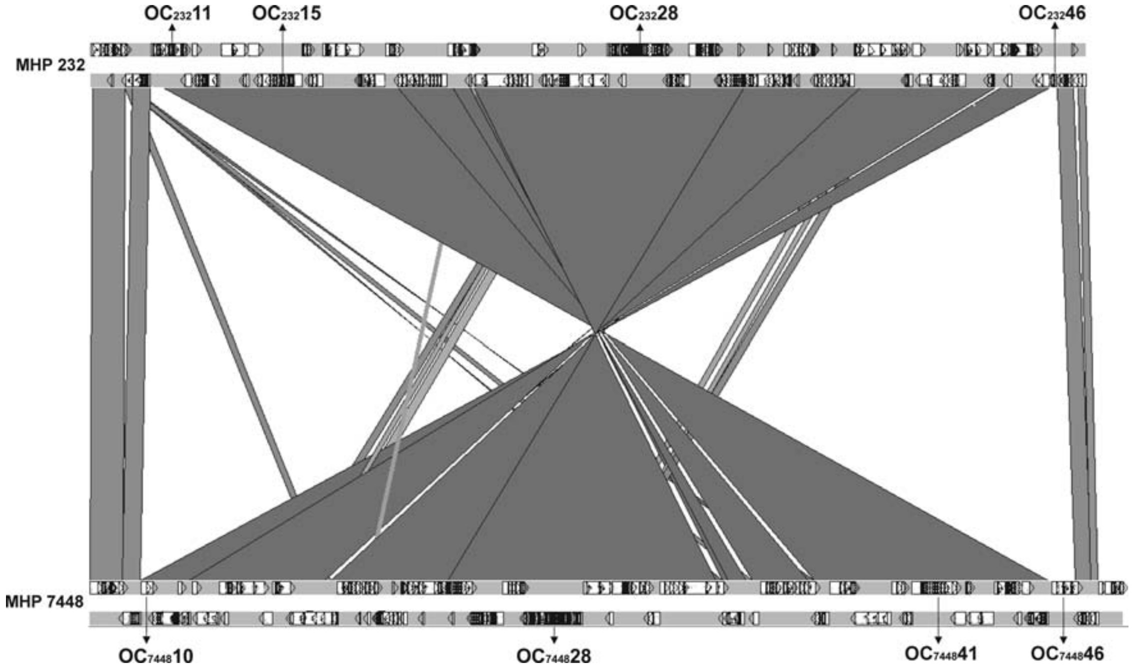

**Figure 3.** A schematic diagram of the inverted region and rearrangement from *M. hyopneumoniae* 232 and 7448 strain. The bars connecting the maps represent close matches. The genomic limits of these regions are shown by arrows ($OC_{232}11$, $OC_{232}46$, $OC_{7448}10$ and $OC_{7448}46$). The other detached OCs are discussed in the text. Visualization by the software Artemis Compare.

comparative genomics is a plausible approach for predicting OCs in mycoplasma genomes.

Some of the differences in ORF strings found among the three *M. hyopneumoniae* strains could be related to the presence of transposases or to the integrative conjugative element.[3] Although the number of transposases is almost similar in all the three genomes (Table 1), their localization in some cases, probably, could be involved in the generation of instable regions where most changes occur, leaving the rest of the genome stable. The integrative conjugative element (ICEH), related to genome rearrangement and probably pathogenesis, is present in the genomes of *M. hyopneumoniae* 7448 and 232 strains but have a different organization.[3] The ICEH is located within $OC_{7448}70$, together with ORFs usually associated with bacterial conjugative elements and ORFs encoding hypothetical products, and within $OC_{232}86$ with a similar organization.

### 3.3. *Experimental analysis of OCs cotranscription*

Aiming to validate some of the OCs predicted by our *in silico* analysis, 21 OCs encompassing 115 ORFs from *M. hyopneumoniae* 7448 were further characterized, corresponding to 17% of the total ORFs in the genome and 18% of the predicted OCs (Table 1 and Supplementary Table S6). The selection of these OCs was initially based on functional category variation among the ORFs and OCs with ORFs with long IRs. Usually, experimental validation of TUs is often established by measuring the length of mRNAs through Northern analysis. However, this approach failed to yield reliable transcript sizes in our experiments (data not shown). This experimental difficulty was also reported for other mycoplasmas and is probably due to the large size and instability of the transcripts in this genus.[7,8] Therefore, we performed gene-specific RT–PCR strategy to amplify the IRs between the ORFs of the predicted $OC_{7448}$ for experimental validation. Briefly, as shown in Fig. 1, we designed primer pairs that would only amplify a product if adjacent genes could be found on a single mRNA molecule. In the example shown in Fig. 1, $OC_{7448}111$ encompassed four ORFs and one tRNA gene and amplicons were generated from: MHP7448_0653 and *ksgA* (amplicon 263 bp), *ksgA* and *rpsL* (amplicon 706 bp), *rpsL* and tRNA30 (amplicon 357 bp), MHP7448_0653 and *rpsL* (amplicon 1555 bp), *ksgA* and tRNA30 (amplicon 1157 bp), indicating that a polycistronic mRNA spanning from MHP7448_0653 to tRNA30 is produced.

The experimental data from RT–PCR analysis of *M. hyopneumoniae* 7448, using primer pairs to each gene spanning the junctions between genes, demonstrated that the ORFs localized within each OC were linked. Therefore, 18% of the OCs characterized *in silico* in the 7448 strain were validated, corresponding to 21 of 117 OCs (Supplementary Table S6). To better demonstrate these results, three segments of the *M. hyopneumoniae* 7448 genome encompassing

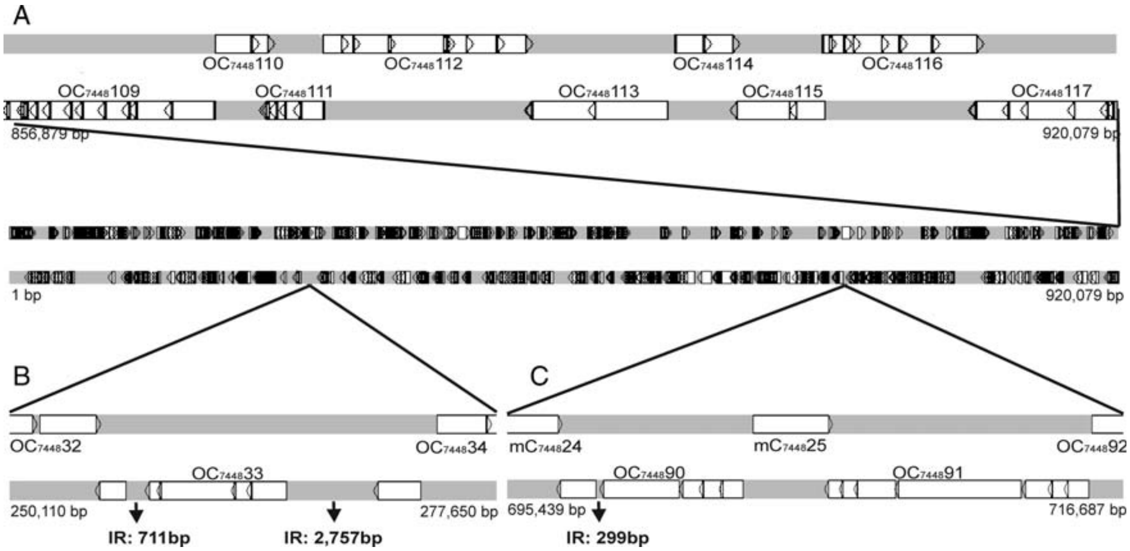

**Figure 4.** Representation of experimentally analyzed OCs in *M. hyopneumoniae* genome. Some regions experimentally tested (**A**–**C**) are detached from the overall picture of the genome (labelled from 1 to 920 079 bp). The position on the chromosomal sequence is indicated in bp below both termini of the bars, with white arrowheads indicating the direction of transcription. (**B** and **C**) The size of some intergenic regions are marked (IR: followed by the number). (**A**) Genome coverage of 63.2 kb showing nine OCs. (**B**) Genome coverage of 27.5 kb showing three OCs. Two intergenic regions (IR: 711 bp and IR: 2757 bp) are indicated. (**C**) Genome coverage of 21.2 kb showing three OCs and two mC. The intergenic region (IR: 299 bp) in OC$_{7448}$90 represent the sequence between *tuf* and *lon* genes. Visualization by the software Artemis.

63.2 kb (Fig. 4A), 27.5 kb (Fig. 4B) and 21.4 kb (Fig. 4C) are shown. The first genome region analysed contained nine OCs (OC$_{7448}$109–OC$_{7448}$117) with a total of 44 ORFs (Fig. 4A), representing the typical ORF string distribution as previously described: (i) ORF number: two ORFs in OC$_{7448}$110, OC$_{7448}$113, OC$_{7448}$114, OC$_{7448}$115 up until 11 ORFs in OC$_{7448}$109; (ii) ORF product: OC$_{7448}$114 products from the same functional category and OC$_{7448}$112 products not functionally related and (iii) ORF string conservation: OC$_{7448}$112 with conserved gene order in OC$_{J}$112 and OC$_{232}$111; OC$_{7448}$113 with same gene order as OC$_{J}$113 and OC$_{232}$112 (Supplementary Tables S4–S6).

Some reports have suggested that genes co-transcribed as TUs or operons are likely to be compactly arranged on the genome with short IRs, although there are cases where regulatory elements can cause variations in the length of TUs.[30,31] *Mycoplasma hyopneumoniae* has a reduced genome with limited biosynthetic potential, and the presence of short intergenic sequences would help us to allow efficient transcription. Interestingly, the size of the IRs was extremely variable within different OCs (Supplementary Tables S3–S5). RT–PCRs with specific primers were applied to amplify all the IRs present in the 21 OCs and cotranscription was observed, even with large IRs (Fig. 4B, Supplementary Table S6). The OC$_{7448}$33 contained six ORFs (Supplementary Table S6) with a significant variation in the IR: (i) IR length of 2757 bp between *glyA* and *nrdF* genes; (ii) IR length of 2 bp between *nrdF* and *nrdI* genes; (iii) IR length of 76 bp between *nrdI* and *nrdE* genes; (iv) IR length of 89 bp between *nrdE* gene and an ORF encoding a hypothetical protein (MHP7448_0220) and (v) IR length of 711 bp between two ORFs encoding hypothetical proteins (MHP7448_0220 and MHP7448_0219).

### 3.4. Cotranscription of ORF strings in the same DNA strand

The genome region detailed in Fig. 4C illustrates the analysis to validate the criteria used *in silico* for OC definition, based on the presence of ORF strings in the same DNA strand until the occurrence of ORFs in the opposite strand. The ORF strings in OC$_{7448}$90 and OC$_{7448}$91 were interrupted by the presence of mC$_{7448}$25 in the opposite strand. For the RT–PCRs, two primer pairs were designed to evaluate the presence of a polycistronic transcript corresponding to the ORFs strings in OC$_{7448}$90 plus OC$_{7448}$91. The first primer pair (gyrF/gyrR, Supplementary Table S1) amplified a 941 bp DNA fragment of the MHP7448_0528_gyrA (mC$_{7448}$25). The other primer pair was localized within the coding sequence of first ORF (*deo*) of OC$_{7448}$90 and in the IR between OC$_{7448}$90 and OC$_{7448}$91 (gyr/deoC, Supplementary Table S1), potentially generating an amplification of 1338 bp. Although the presence of mRNA for the *gyrA* ORF was demonstrated, no transcript (single polycistronic mRNA) was observed, including both OC ($OC_{7448}90$ plus $OC_{7448}91$) ORF strings. Therefore, these results support the criteria used *in silico* for defining the OCs characterized by the presence of two or more ORFs in the same DNA strand until the occurrence of ORFs in the opposite strand. Taken together, these findings demonstrated that the typical gene organization present in other bacteria, as *Escherichia coli* or *Bacillus subtilis*, cannot be applied to *M. hyopneumoniae* genome.

Although in *M. hyopneumoniae* genes are arranged in long clusters, suggesting that the starting point of transcription occurs in the first gene, the transcription termination regions have yet to be defined. Aiming to contribute to the definition of the terminator sequences, one IR between $OC_{7448}112$ and $OC_{7448}113$ (see Fig. 1A) was analysed by RT–PCR. Primers pairs were designed to amplify the region between the ends of each OC (Supplementary Fig. S1). Our results demonstrate the presence of transcripts in this region, suggesting the absence of specific regions for transcription termination. These findings are supported by the observation of Gardner and Minion[18] that transcription occurs in IRs of *M. hyopneumoniae*, and that transcription run off the end of the last ORF.

The gene encoding the elongation factor Tu (*tuf*) is the last of five ORFs to constitute an OC ($OC_{7448}90$), expressed from a promoter upstream of the first ORF (*deoC*) (Supplementary Table S6, Fig. 4C). However, in the 299 bp IR between *tuf* and *lon*, a promoter-like sequence was found by primer extension analysis (data not shown). The $OC_{7448}90$ represents a typical OC in *M. hyopneumoniae*, with ORFs encoding products not related functionally as *deoC* (deoxyribose-phosphate aldolase), *upp* (uracil phosphoribosyltransferase), MHP7448_0525 (hypothetical protein), *lon* (heat shock ATP-dependent protease) and *tuf* (elongation factor Tu, EF-Tu).

EF-Tu is necessary for protein synthesis in metabolically active cells, being possibly involved in the synthesis of alternative transcripts in the *M. hyopneumoniae* 7448 strain. It is already known that, in TUs, discoordinate expression occurs when the product of an individual gene needs to be expressed apart from the others to mediate a different role.[31,32] The $OC_{7448}90$ has upstream and internal promoters resulting in two different sizes of clusters, indicating that the entire OC could be expressed from a common promoter or a subset of the OC (*tuf*) is separately transcribed in response to a specific signal. Previous reports have suggested a relationship between the expanded IRs and presence of internal regulatory elements in highly expressed TUs.[31,33] Güell *et al.*[12] in an *M. pneumoniae* transcriptome analysis demonstrated many alternative transcripts. The experimental evidence of the presence of regulatory elements such as internal promoters in the expanded IRs provided by the analysis of the *tuf* gene (IR of 299 bp between adjacent genes) and the variable size of the *M. hyopneumoniae* IRs analysed by RT–PCR suggest the presence of internal promoters in the IRs of *M. hyopneumoniae* OCs and a complexity in transcription regulation.

## 4. Conclusions

Different approaches have been used to analyse gene organization in bacterial genomes through the development of new algorithm[34] or prediction programs,[35,36] which are based on characteristics such as intergenic distance, closely related gene function, conservation of gene order in phylogenetically close genomes, or even the prediction of promoter region[37] and transcriptional terminators.[19,20] Although these algorithms have shown a great promise in terms of being able to predict operon structure with a high degree of specificity and sensitivity, the data that they rely on are available only for a small number of bacterial species[36] and their application to all bacterial species could present some difficulties. Furthermore, combinations of these prediction analyses with experimental data are scarce. Not surprisingly, analysis of OC prediction followed by experimental confirmation in mycoplasma genomes was not conducted prior to this study. Moreover, experimental data on TUs are also scarce.

Similar to *M. hyopneumoniae* genes, some *M. pneumoniae* genes are arranged in long OCs but only with short or almost no intergenic sequences.[12] Analysis of P97 and P102 paralog families have also suggested that a distance of up to 54 bp is an upper limit for defining the cotranscript structure in *M. hyopneumoniae*.[9] However, we have systematically mapped *in silico* all the OCs in three different strains of *M. hyopneumoniae* and experimentally validated 115 ORFs in 21 OCs with variably sized IRs in the 7448 strain. It has been proposed that intergenic distances between ORFs in OCs are similar in prokaryotes and can be used to predict OCs and to estimate the total number of OCs in genomes.[26,28,29,38,39] However, according to our analysis, the distance models do not seem to be applicable for OC prediction in *M. hyopneumoniae* genomes. Furthermore, the variability in size of intergenic distances within an OC can be related to the complexity of mechanisms of gene regulation. Therefore, the unusual distribution of intergenic distances between ORFs within OCs in *M. hyopneumoniae* may reflect a biological difference in the genome structure as previous suggested for *Synechocystis*.[28] Furthermore, transcription analysis across the IRs of the *M. hyopneumoniae* 232 genome demonstrated the presence of transcripts in the IRs and that transcription run off the end of the last ORF and RNA polymerase is gradually released from template.[18]

Taken together, these findings suggest that OCs are continuously transcribed until reaching the next OC or ORF present in the opposite strand. However, a subset of the OCs could be independently transcribed due to the presence of internal promoters. These new data suggest complex transcriptional organization in *M. hyopneumoniae* genomes and the probability of the occurrence of other not-yet-known mechanisms that are involved in transcription. Furthermore, the high synteny of gene organization between pathogenic and non-pathogenic strains indicates that pathogenicity is not related to gene organization. The accelerating pace of molecular research into this important pathogen is certain to provide additional data to refine the results described in this work, forming a solid empirical foundation for our future understanding of the relationships between gene organization and transcription regulation in *M. hyopneumoniae*.

**Supplementary data:** Supplementary data are available at www.dnaresearch.oxfordjournals.org.

## Funding

This work was supported by grants from the Brazilian National Research Council (CNPq).

## References

1. Ross, R.F. 1992, Mycoplasmal disease. In: Leman, A.D., Straw, B.E., Mengeling, W.L., D'Allaire, S., and Taylor, D.J. (eds). *Diseases of Swine*. Iowa State University Press: Ames, pp. 537–51.
2. Minion, F.C., Lefkowitz, E.L., Madsen, M.L., et al. 2004, The genome sequence of strain 232, the agent of swine mycoplasmosis, *J. Bacteriol.*, **186**, 7123–33.
3. Vasconcelos, A.T., Ferreira, H.B., Bizarro, C.V., et al. 2005, Swine and poultry pathogens: the complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae*, *J. Bacteriol.*, **187**, 5568–77.
4. Liu, W., Feng, Z., Fang, L., et al. 2011, Complete genome sequence of *Mycoplasma hyopneumoniae* strain 168, *J. Bacteriol.*, **193**, 1016–7.
5. Inamine, J.M., Loechel, S., and Hu, P.C. 1988, Analysis of the nucleotide sequence of the P1 operon of *Mycoplasma pneumoniae*, *Gene*, **73**, 175–83.
6. Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. and Herrmann, R. 1997, Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*, *Nucl. Acids Res.*, **25**, 701–12.
7. Waldo, R.H. III, Popham, P.L., Romero-Arroyo, C.E., Mothershed, E.A., Lee, K.K., and Krause, D.C. 1999, Transcriptional analysis of the *hmw* gene cluster of *Mycoplasma pneumoniae*, *J. Bacteriol.*, **181**, 4978–85.
8. Musatovova, O., Dhandayuthapani, S., and Baseman, J.B. 2003, Transcriptional starts for cytadherence-related operons of *Mycoplasma genitalium*, *FEMS Microbiol. Lett.*, **229**, 73–81.
9. Adams, C., Pitzer, J., and Minion, F.C. 2005, *In vivo* expression analysis of the P97 and P102 paralog families of *Mycoplasma hyopneumoniae*, *Infect. Immun.*, **73**, 7784–7.
10. Benders, G.A., Powell, B.C., and Hutchison Iii, C.A. 2005, Transcriptional analysis of the conserved *fts*Z gene cluster in *Mycoplasma genitalium* and *Mycoplasma pneumoniae*, *J. Bacteriol.*, **187**, 4542–51.
11. Waldo, R.H. III and Krause, D.C. 2006, Synthesis, stability, and function of cytadhesin P1 and accessory protein B/C complex of *Mycoplasma pneumoniae*, *J. Bacteriol.*, **188**, 569–75.
12. Güell, M., Noort, V., Yus, E., et al. 2009, Transcriptome complexity in a genome-reduced bacterium, *Science*, **326**, 1268–71.
13. Muto, A., and Osawa, S. 1987, The guanine and cytosine content of genomic DNA and bacterial evolution, *Proc. Natl. Acad. Sci. USA*, **84**, 166–9.
14. Weiner, J. III, Herrmann, R., and Browning, G.F. 2000, Transcription in *Mycoplasma pneumoniae*, *Nucl. Acids Res.*, **28**, 4488–96.
15. Dybvig, K., French, C.T., and Voelker, L.L. 2000, Construction and use of derivatives of transposon Tn4001 that function in *Mycoplasma pulmonis* and *Mycoplasma arthritidis*, *J. Bacteriol.*, **182**, 4343–7.
16. Halbedel, S. and Stulke, J. 2006, Probing *in vivo* promoter activities in *Mycoplasma pneumoniae*: a system for generation of single-copy reporter constructs, *Appl. Environ. Microbiol.*, **72**, 1696–9.
17. Janis, C., Lartigue, C., Frey, J., et al. 2005, Versatile use of *oriC* plasmids for functional genomics of *Mycoplasma capricolum* subsp. capricolum, *Appl. Environ. Microbiol.*, **71**, 2888–93.
18. Gardner, S.W., and Minion, F.C. 2010, Detection and Quantification of intergenic transcription in *Mycoplasma hyopneumoniae*, *Microbiology*, **156**, 2305–15.
19. Washio, T., Sasayama, J., and Tomita, M. 1998, Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination, *Nucl. Acids Res.*, **26**, 5456–63.
20. Hoon, M.J.L., Makita, Y., Nakai, K., and Miyano, S. 2005, Prediction of transcriptional terminators in *Bacillus subtilis* and related species, *PLoS Comput. Biol.*, **1**, 212–21.
21. Madeira, H.M.F., and Gabriel, J.E. 2007, Regulation of gene expression in mycoplasmas: contribution from *Mycoplasma hyopneumoniae* and *Mycoplasma synoviae* genome sequences, *Genet. Mol. Biol.*, **30**, 277–82.
22. Rutherford, K., Parkhill, J., Crook, J., et al. 2000, Artemis: sequence visualization and annotation, *Bioinformatics*, **16**, 944–5.

23. Carver, T., Berriman, M., Tivey, A., et al. 2008, Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database, *Bioinformatics*, **24**, 2672–6.
24. Friis, N.F. 1975, Some recommendations concerning primary isolation of *Mycoplasma suipneumoniae* and *Mycoplasma flocculare* a survey, *Nordisk Veterinaer. Medicin.*, **27**, 337–9.
25. Sambrook, J., and Russell, D.W. 2001, *Molecular Cloning a Laboratory Manual*. Cold Spring Harbor Laboratory Press: New York.
26. Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., and Wolf, Y.I. 2002, Connected gene neighborhoods in prokaryotic genomes, *Nucl. Acids Res.*, **30**, 2212–23.
27. Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y., and Jianq, T. 2004, Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome, *Nucl. Acids Res.*, **32**, 2147–57.
28. Price, M.N., Huang, K.H., Arkin, A.P., and Alm, E.J. 2005, A novel method for accurate operon predictions in all sequenced prokaryotes, *Nucl. Acids Res.*, **33**, 880–92.
29. Roback, P., Beard, J., and Baumann, D. 2007, A predicted operon map for mycobacterium tuberculosis, *Nucl. Acids Res.*, **35**, 5085–95.
30. Cho, B.K., Zengler, K., Qiu, Y., et al. 2009, The transcription unit architecture of the *Escherichia coli* genome, *Nat. Biotechnol.*, **27**, 1043–9.
31. Okuda, S., Kawashima, S., Kobayashi, K., Ogasawara, N., Kanehisa, M., and Goto, S. 2007, Characterization of relationships between transcriptional units and operon structures in *Bacillus subtilis* and *Escherichia coli*, *BMC Genomics*, **8**, 48.
32. Adhya, S. 2003, Suboperonic regulatory signals, *Sci. STKE*, **185**, pe22, 1–7.
33. Price, M.N., Arkin, A.P., and Alm, E.J. 2006, The life-cycle of operons, *PLoS Genet.*, **2**, 859–73.
34. Yang, Q. and Sze, S.H. 2008, Large-scale analysis of gene clustering in bacteria, *Genome Res.*, **15**, 949–56.
35. Mao, F., Dam, P., Chou, J., Olman, V. and Xu, Y. 2009, DOOR: a database for prokaryotic operons, *Nucl. Acids Res.*, **37**, 459–63.
36. Bergman, N.H., Passalacqua, K.D., Hanna, P.C., and Qin, Z.S. 2007, Operon prediction for sequenced bacterial genomes without experimental information, *Appl. Environ. Microbiol.*, **73**, 846–54.
37. Jian-Cheng, L., Jin-Lin, X., Jian-Hua, L., and Yi-Xue, L. 2003, Prediction of prokaryotic promoters based on prediction of transcriptional units, *Acta Biochimica et Biophysica Sinica*, **35**, 317–24.
38. Moreno-Hagelsieb, G., and Collado-Vides, J. 2002, A powerful nonhomology method for the prediction of operons in prokaryotes, *Bioinformatics*, **18**, 329–36.
39. Brouwer, R.W.W., Kuipers, O.P., and Van Hijum, S.A.F.T. 2008, The relative value of operon predictions, *Brief Bioinform.*, **9**, 367–75.