Eukaryotic hosts could regulate the gene expression of the endosymbiont bacterium *Wolbachia pipientis*

Stella Papaleo[1$], Simona Panelli[1], Ibrahim Bitar[2], Lodovico Sterzi[1], Riccardo Nodari[1,3], Francesco Comandatore[1*$]

1. Department of Biomedical and Clinical Sciences, Pediatric Clinical Research Center "Romeo and Enrica Invernizzi", University of Milan, 20157 Milan, Italy.
2. Biomedical Center, Faculty of Medicine, Charles University, Pilsen, Czechia.
3. Istituto Nazionale di Genetica Molecolare (INGM) "Romeo and Enrica Invernizzi", University of Milan, 20157 Milan, Italy.

* Corresponding author

$ The authors contributed equally

Wolbachia pipientis is an important endosymbiont bacterium in arthropods and filarial nematodes, being essential for survival and reproduction of the hosts. The evolution of such a strict symbiotic relationship led to the erosion of a large part of the Wolbachia genome, including most gene regulatory regions. Despite this, experimental studies show Wolbachia gene expression varies during host development and/or after environmental stresses, but the mechanism is still not clear. The first published RNA-Seq study on Wolbachia evidenced the differential expression of a putative DNA methyltransferase, an enzyme able to methylate adenine or cytosine on specific patterns and often involved in the regulation of bacterial gene expression. In this work, we tested the hypothesis that the activity of this enzyme modulates the gene expression in Wolbachia. We expressed the Wolbachia DNA

methyltransferase gene in *Escherichia coli* to test its biological activity and to determine its methylation pattern; then we re-analysed RNA-Seq data from six studies. We surprisingly found that the Wolbachia DNA methyltransferase is the only one described so far capable of methylating both adenine and cytosine without recognising highly conserved patterns. We also found that the nucleotide content of Wolbachia genes correlates with their expressions, with a pattern compatible to be a consequence of DNA methylation. Lastly, we found a conserved eukaryotic-like promoter upstream of the Wolbachia DNA methyltransferase gene. Overall, these findings suggest a cascade mechanism in which the host induces the Wolbachia DNA methyltransferase expression, which in turn modulates the bacterium gene expression.

Introduction

Wolbachia pipientis (from here Wolbachia) is probably one of the most interesting and investigated endosymbiont bacteria, colonising several species of arthropods and filarial nematodes[1,23]. In arthropods, the bacterium can manipulate the reproduction of the host[4] , protect it from viruses[5,6] and provide nutrients[7]. In filarial nematodes, genomic analyses and experimental evidence revealed that Wolbachia is essential for the host survival[8]. The bacterium has typically an intracellular tropism, often colonising microtubules and endoplasmic reticulum (ER) with important effects on host cell replication[3]. It is mainly mother-to-offspring transmitted, being actively transported by specialised cells of the host from the soma to germ cells [3]. Nevertheless, horizontal transfer among host individuals can also occur [9,10] and are important for the bacterium spreading [10].

As for many other strictly host-associated bacteria, Wolbachia genome underwent an important reduction during its evolution, reaching a size that ranges from ~0.8 Mb to ~1.8 Mb (BV-BRC representative genomes)[11]. As observed for several endosymbiont bacterial species, genome reduction passes through the increasing of the number of mobile elements (e.g. Insertion Sequences) and the loss of unnecessary genes[12–14]. During this process, the structure of the genome and the order of genes can change dramatically[13], leading to the disruption of several operons and loss of most gene promoters. Interestingly, RNA-seq studies have highlighted that Wolbachia gene expression can be influenced by host environmental/developmental conditions[15–21]. This implies that Wolbachia likely possesses a mechanism for regulating gene expression, even though the low number of promoters indicates that it may not rely solely on transcription factors.

The first published Wolbachia RNA-seq study[15] compared the gene expression of Wolbachia isolated from the soma and gonad in females and males of the filarial nematode *Onchocerca ochengi*. One of the strongest differential expressions was reported for a gene annotated as

a S-adenosyl-methyonine-dependent (SAM)-dependent DNA methyltransferase (from here MTase). In general, DNA methyltransferase enzymes are able to methylate adenine or cytosine on specific sequence patterns of bacterial DNA. For instance, the DNA adenine methyltransferase (Dam) methylates the adenine residues in the sequence GATC, and the DNA cytosine methyltransferase (Dcm) methylates the second cytosine of the sequences CCAGG and CCTGG [22,23]. Despite DNA methylation being a well understood mechanism of gene expression regulation in eukaryotes (i.e. epigenetics), its role in prokaryotes physiology is still poorly investigated and no studies have been published on Wolbachia. At the state of the art, there is evidence that DNA methyltransferase enzymes are involved in fundamental bacterial processes, including proofreading during DNA replication, protection from foreign DNA and regulation of gene expression. A specific type of DNA methyltransferase, called orphan DNA methyltransferases, methylates gene promoters, changing their affinity for transcription factors with important effects on gene expression [22,23].

In this study, we investigated the hypothesis that DNA methylation is a key mechanism for the regulation of the Wolbachia gene expression. We experimentally determined the methylation pattern of the MTase enzyme encoded in the genome of the Wolbachia endosymbiont of *O. ochengi* (wOo). Then, we tested whether the frequency of this pattern into gene sequences correlates with their expression, and whether the intensity of the phenomenon is related to the expression of the MTase gene. Finally, we identified a highly conserved putative eukaryotic-like promoter upstream of the MTase gene on several Wolbacha genomes.

Our results show that Wolbachia harbours a peculiar SAM-dependent DNA methyltransferase (MTase) enzyme and that this enzyme could regulate Wolbachia gene expression, and that the expression of this MTase could be in turn regulated by Wolbachia hosts.

Material and methods

Experimental determination of the wOo SAM-dependent DNA methyltransferase methylation pattern

The wOo SAM-dependent DNA methyltransferase (MTase) was artificially synthesised with codon usage adaptation and cloned into the XhoI and SalI sites of the pCOLD III expression vector (Takara Bio). The recombinant construct was transformed into *Escherichia coli* StellarTM Competent Cells (*dam*–/*dcm*–) (Takara Bio) according to the supplier's protocol (ClonTech; Protocol-at-a-Glance, PT5056-2). The *E. coli* strains were cultured in LB broth medium supplemented with ampicillin. Before inducing expression, PCR with custom primers targeting the gene of interest was performed on bacterial crude extracts to verify the outcome of transformation. MTase expression was then induced at 15°C, according to the protocol provided by the manufacturer for the Cold Shock Expression System of pColdTM plasmids (Takara Bio). 24 hours after the induction, the expression of the wOo MTase gene was verified by SDS-PAGE. DNA was then extracted using the NucleoSpin Microbial DNA Mini kit (Macherey Nagel) and subsequently subjected to long reads sequencing using Sequel I (Pacific biosciences). Assembly of reads was done using the Microbial Assembly pipeline provided by SMRT Link v.10.1 with a minimum seed coverage of 30x. The assembled genome was used as a reference in the downstream analysis. The methylation status of each base of the obtained reads was then determined using the 'Base modification Analysis' pipeline included in SMRT-Portal. The pipeline performs base modification and modified base motifs detection. A control experiment was carried out transforming *E. coli* Stellar TM Competent Cells with a pCOLD III vector without the wOo MTase gene inserted. The 95th percentile of Mean QV and coverage values obtained from the control experiment were set as minimum thresholds for the identification of methylated bases on the transformed *E. coli* strain. These last analyses were performed using R.

Dataset reconstruction and data normalisation

The table S2 and gff map files of the supplementary information of Chung et al. 2020 were retrieved [24]. The table contains the gene expression quantifications computed by Chung and colleagues (2020) re-analysing seven RNA-Seq studies on Wolbachia [15–21], for a total of 128 samples (replicates) from 62 host conditions. More in details: i) four samples from two conditions on Wolbachia endosymbiont of *Onchocerca ochengi* (wOo) from Darby et al. 2012[15]; ii) five samples from five conditions on Wolbachia endosymbiont of *Dirofilaria immitis* (wDi) from Luck et al., 2014[16]; iii) six samples from two conditions on Wolbachia endosymbiont of *Drosophila melanogaster* (wMel), from Darby et al., 2014[17]; iv) 12 samples from seven conditions on wDi from Luck et al., 2015[18]; v) 55 samples from 24 conditions on wMel, from Gutzwiller et al., 2015[19]; vi) 14 samples from seven conditions on Wolbachia endosymbiont of *Brugia malayi* (wBm) from Grote et al., 2017[20]; vii) 32 samples from 15 conditions on wBm from Chung et al., 2019[21]. Preliminarily, we excluded from the dataset all the samples from Luck et al., 2015[18] because only 8% of the gene expression quantifications were greater than 0. Thus, gene expression analysis was performed on a total of 116 samples from 55 host conditions, two for wOo, five for wDi, 22 for wBm and 26 for wMel.

Gene expression values were normalised to make them comparable among studies and conditions. For each condition in each study the gene expression values were expressed as

$$\text{Log2expr} = \sum_s [\sum_i \log2(\text{expr}_{s,i} / \sum_i \text{expr}_{s,i})] / S_{num}$$

Where:

s = sample (or replicate); i = gene; $\text{expr}_{s,i}$ = the gene expression value of the gene i in the sample (replicate) s; S_{num} = total number of samples. This normalisation formula derives from the log2 fold change formula often used to compare two conditions in RNA-Seq studies:

$$\text{Log2FC} = \text{Log2expr}_{condition1} - \text{Log2expr}_{conditio21} =$$

$$\{\sum_s [\sum_i \log2(\text{expr}_{s,i} / \sum_i \text{expr}_{s,i})] / S\}_{condition1} - \{\sum_s [\sum_i \log2(\text{expr}_{s,i} / \sum_i \text{expr}_{s,i})] / S\}_{condition2}$$

Log2FC allows us to compare values from different conditions in the same experiment. To make the normalised values more comparable among the studies, they were converted in percentiles. Indeed, the percentile may be less biassed by sample specific features, such as the number of genes.

Genes composition parameters determination

The genome assemblies of wBm (AE017321.1), wDi (www.nematodes.org: wDi 2.2), wOo (HE660029.1) and wMel (AE017196.1) were retrieved and passed to Prodigal [25] for Open Reading Frame (ORF) calling. For each ORF, Codon Bias Index (CBI) (which measure how much a gene uses a subset of optimal codons, ranging from 0 for random using, to 1 for maximum bias), the effective number of codons (Nc) and the ORF length were computed using CodonW tool (Peden J.F. Analysis of codon usage. 1999. PhD Thesis, University of Nottingham, UK), while %A, %T, %G, %C and %G/%T (from here G/T ratio) were computed using an in-house Perl script.

Statistical analyses

The relationship between each gene composition parameter (CBI, Nc, length, %A, %T, %G, %C and G/T ratio, see above) and gene expression was investigated. For each of the 55 samples included in the study, linear regression was used to assess the correlation between gene expression and each of the gene composition parameters. For each regression line, the slope and the p value were obtained. Slope values report the rate of change of gene expression in relationship to the gene composition parameter: the higher the slope, the more the gene expression varies as a consequence of the gene composition parameter. Thus, to investigate if MTase activity affects the expression of genes on the basis of their composition parameters, the linear regression between the MTase gene expression and regression line slope were computed for each gene composition parameters.

The presence of other genes with expression patterns comparable to that of MTase gene was tested as follows. Exploiting the gene orthologous information included in the Table S6 of Chung et al., 2020 [24], we performed a Spearman co-correlation among the expressions of the 546 single copy core genes shared among wOo, wDi, wBm and wMel. The expression patterns were also investigated by Principal Component Analysis (PCA), using R.

Comparison between species tree and MTase tree

A dataset including 112 high quality Wolbachia genome assemblies spanning the host genetic diversity was reconstructed, including 111 assemblies retrieved from the Genome Taxonomy DataBase (GTDB) and the wDi genome used for the previous analysis (indeed, the wBm, wMel and wOo assemblies were already present in the GTDB database) (see Table S1). The genome assemblies were subjected to Prodigal [25] for ORF calling, and the amino acid sequences were passed to OrthoFinder [26] for orthologue analysis. The nucleotide sequences of the single copy core genes were then retrieved, aligned using the MUSCLE tool [27], tested for recombinations using the PhiPack tool [28], trimmed using the trimal tool (-gt 0.5 setting) [29], and finally concatenated. The obtained concatenate was passed to RAxML8 [30] for phylogenetic analysis using the GTR+I+G model, as previously determined using ModelTest-NG tool [31]. The orthologue group containing the MTase genes was determined and subjected to phylogenetic analysis following the same flow described above (using the TVM+I+G model and excluding the step of the concatenation of the genes). The obtained trees were then compared using the Cophylo R library [32].

Analysis of the region upstream the methyltransferase gene

The positions and orientation of the MTase genes on the 112 genome assemblies were determined by BLASTn searches. Then, for each genome assembly, the 200 nucleotides upstream of the MTase gene were extracted and screened for the presence of TATA box (pattern TATAWAWR), CAAT box (pattern HYYRRCCAWWSR) and GC box (WRDRGGHRKDKYYK) as described by Bucher [33], using an in-house Perl scripts. The

sequences were also tested for containing eukaryotic promoters using the Neural Network Promoter Prediction tool 34 available on the Berkeley Drosophila Genome Project website (https://www.fruitfly.org/). Lastly, the sequences were visualised using the gplots R library.

Analysis of the genes with high and low G/T ratio

One of the gene composition parameters found to be more correlated to gene expression and MTase action was G/T ratio. Thus, the expression of the genes with higher or lower G/T ratio is expected to be more affected by the action of the MTase enzyme. To investigate the effects of this mechanism on the Wolbachia metabolism and physiology, genes with low or high G/T ratio values were studied as follows. All the amino acid sequences from the 112 Wolbachia strains included in the study were annotated on the basis of the Clusters of Orthologous Genes (COG) database. For each strain, among the COG-annotated genes, the 10 with lower G/T ratio ("low G/T ratio genes") and the 10 with higher G/T ratio ("high G/T ratio genes") were retrieved. The pattern of presence/absence and G/T ratio value of all these retrieved genes was graphically investigated by producing a heatmap using the gplots R library.

Results

Experimental determination of the wOo SAM-dependent DNA methyltransferase methylation pattern

The expression of the wOo MTase into the Stellar *E. coli dam-/dcm-* strain followed by the SMRT PacBio sequencing allowed to identify 166 high quality methylated bases: 151 on cytosine (classified as m4C - N4-methylcytosine) and 15 on adenine (classified as m6A - N6-methyladenine) (see Table S2 and Figure S1). Investigation of the methylation pattern revealed that m4c methylations occurred in a pattern that was not well conserved, whereas m6A methylation involved a guanine-rich region at positions -4, -8 and -19 (Figure 1).

Dataset reconstruction

For each of the six datasets included in the study, the normalised gene expression values were obtained as described in material and methods. The ORF sequences of the wOo, wDi, wBm and wMel genome assemblies were analysed to determine the CBI, Nc, length, %A, %T, %G, %C and %G / %T (from here G/T ratio). The obtained information about gene composition parameters was merged to relative gene expression data (Supplementary information available on repository...).

Statistical analyses

The slope and p value of the linear regression analyses performed between the seven gene composition parameters (CBI, Nc, gene length, %A, %T, %G, %C and G/T ratio) and gene expression are reported in Table S3. Length and %G significantly correlate with gene expression in all the 55 samples included in the study (100%), while %A in 53/55 (96%), CBI 42/55 (78%), %T 40/55 (73%), G/T ratio 31/55 (56%), %C 22/55 (40%) and Nc 7/55 (13%). The obtained slopes for the regression lines range as follows: for gene length from 0.009 to 0.029, %G from 0.79 to 5.3, CBI from 30.9 to 71.7, %T from -1.8 to 1.1, G/T ratio from 18.5 to 90.6, %C from -1.3 to 2.2 and Nc from 0.2 to 0.8. For each of the seven gene composition parameters, the correlation between the expression of the MTase gene and the regression slope was tested by linear regression. The obtained correlations were significant for all parameters except %C, which had a p-value of 0.08. For %G, %C, %A, G/T ratio, Nc, and length, the slope increases with higher DNA MTase expression. Conversely, %T and CBI exhibit a decreasing trend with increasing of the MTase expression (Figures 2-3 and Figures S2-S7).

Co-correlation analysis among the expression patterns of the single copy core genes shared among wOo, wDi, wBm and wMel showed that no other gene has an expression pattern overlapping that of the DNA MTase gene. Indeed, the highest correlation value was 0.84 and

the median value was 0.02 (Figure S8). In general, the absence of clusters of co-expressed genes is also evident from PCA analysis (Figure S9).

Comparison between species tree and MTase tree

Orthology analysis identified a total of 3,561 orthologous groups, including 263 single copy core genes. None of the single copy core genes were recombined. After trimming of gene alignments, the obtained concatenate had a length of 223,833 bp. On the other side, the MTase gene alignment was 1,342 bp long. The species tree obtained from the 263 single copy genes and the MTgene tree are reported in Figure S10. The obtained topologies were mainly congruent, with few signs of Horizontal Gene Transfer (HGT) events.

Analysis of the region upstream the methyltransferase gene

The MTase gene was found in all the 112 genome assemblies included in the study. Unfortunately, for one genome assembly (accession GCF_018454445.1, Wolbachia endosymbiont of *Rhagoletis cerasi* - wCer5) the MTase gene was located on the extreme of a contig, and thus it was not possible to extract the 200 bp upstream of the gene transcription initiation site. 104 out of the 111 regions upstream of the MTase gene contained one or more canonical TATA box sequences and 14 showed canonical Pribnow boxes. In 25 out of 111 (23%) strains, at least one CAAT box with 1 mismatch was found, while 64 out of 111 (58%) showed at least one GC box with 1 mismatch (Figure 4). Lastly, the Neural Network Promoter Prediction tool available on the Berkeley Drosophila Genome Project website, confirmed for 100 out of 111 upstream regions the presence of putative eukaryotic promoters (Figure 4).

Analysis of the genes with high and low G/T ratio

A total of 57 low G/T ratio genes and 53 high G/T ratio genes were found. Among the low G/T ratio genes, the most frequent COG categories were “Translation, ribosomal structure and biogenesis” (J) with 13/57 (23%) genes, “Mobilome: prophages, transposons” (X) with

8/57 (14%), “Posttranslational modification, protein turnover, chaperones” (O) with 7/57 (12%) and “Transcription” (K) with 5/57 (9%) genes. Conversely, among the high G/T ratio genes, the most frequent COG categories were “Energy production and conversion” (C) with 15/53 (28%) genes, “Inorganic ion transport and metabolism” (P) with 11/53 (21%), “Lipid transport and metabolism” (I) with 5/53 (9%) genes. The presence/absence and G/T ratio patterns are shown in Figure 5.

Discussion

Wolbachia pipientis is one of the most diffused endosymbiont bacteria [9], essential for the survival and reproduction of several arthropod and filarial nematode host species [1,3,34]. The close symbiotic relationship with the host has had important consequences on the bacterial genome structure, including an important reduction [35], with frequent rearrangements and the loss of non-essential regions [13], which has caused the disruption of several operons and promoters. The loss of promoters could have limited evolutionary costs, considering that the intracellular tropism of the bacterium protects it from external environment stresses. However, experimental studies showed the existence of Wolbachia genes differentially expressed in host samples from different body compartments, developmental stages and/or from hosts subjected to environmental stress [24]. Nevertheless, at the state of the art, the gene regulation mechanisms in Wolbachia are poorly understood.

One of the most differentially expressed genes between Wolbachia colonising the soma and gonads of the filarial nematode *Onchocerca ochengi* (wOo) was a SAM-dependent DNA methyltransferase (MTase) [15]. Methylation of bacterial gene promoters is known to influence the promoters affinity for transcription factors, leading to changes in gene expression and epigenetic modifications [22,36].

In this work, we investigated the role of Wolbachia MTase as a mediator of gene expression regulation. In absence of promoters, we hypothesise that the intra-gene DNA methylation could affect gene expression. To test this hypothesis, we i) experimentally determined the sequence pattern recognised by the wOo DNA MTase; ii) studied expression data of 55 samples from six published RNA-Seq studies on Wolbachia; iii) investigated the presence of putative eukaryotic-like promoters upstream the DNA MTase gene.

In silico determination of the sequence pattern recognized by MTase is challenging [37]. This is especially true in this case, considering that this is the first Wolbachia DNA MTase enzyme ever characterised. An experimental approach to determining the methylation pattern of the enzyme is to express the MTase in an *Escherichia coli* strain depleted for any MTase gene, and then identify the methylated bases and motifs using SMRT PacBio sequencing [37]. Surprisingly, wOo MTase was found to be able to methylate both adenine and cytosine, without highly conserved patterns (Figure 1), especially for cytosine. Our findings also suggest that the enzyme has a 10-fold greater affinity for cytosine than adenine. To our knowledge this is the first MTase enzyme described with these peculiar features.

Regarding the analysis of gene expression, we started from the idea that if DNA methylation within the gene affects its expression, we should observe a correlation between gene nucleotide composition and gene expression. Furthermore, the direction (positive or negative) and the strength (the slope of the regression line) of this correlation should change with variation of DNA MTase gene expression. It is well-documented in literature that gene length and codon usage could be correlated to gene expression [37,38]. In particular, highly expressed genes tend to have lower Effective Number of Codons (Nc) and higher Codon Bias Index (CBI). For each of the 55 samples, we computed the correlation between CBI, Nc, gene length, %A, %T, %G, %C and G/T ratio versus the relative gene expression. The result shows correlation between gene nucleotide composition and gene expression in

several samples. Interestingly, we found that the slope of the obtained regression lines often correlates with the MTase expression in that sample.

In more detail, the expression of genes is positively correlated with their %G in all samples and the correlation slope increases proportionally to the expression of DNA MTase (Figure S6). Regarding %T, the correlation with gene expression was positive (the higher the %T the higher the gene expression) in samples with poorly expressed MTase, while it became negative (the higher the %T the lower the gene expression) in samples with high MTase gene expression (Figure S5). Conventionally, gene sequences are complementary to the transcribed DNA strand, thus corresponding to the messenger RNA sequence. It means that the observed correlation between gene expression and %G and %T may be interpreted as the effects of the methylation of respectively cytosine and adenine on the transcribed strand. The positive to negative slope trend observed for %T is suggestive of the importance that methylation of adenine may have in the regulation of gene expression, in particular considering the AT-richness of the Wolbachia genome. Lastly, we found that the ratio of %G and %T (G/T ratio) is a good parameter to study Wolbachia gene expression: it significantly positively correlates with gene expression in all samples except for those with very low expression of MTase (Figure 2). This strongly supports the hypothesis that the methylations of cytosine and adenine could be the key to understanding the regulation of gene expression in Wolbachia.

Furthermore, it is interesting to note that highly expressed genes have high Codon Bias Index (CBI) (as expected from literature) only for samples with low levels of MTase expression; in those with highly expressed MTase this correlation became not significant (Figure 3). This suggests that the MTase effect on gene expression may be evolved independently from the codon usage adaptation.

The step forward in our analysis was to investigate how the expression of the MTase gene can be regulated. To this aim we analysed a collection of representative Wolbachia

genomes, screening the 200-bp region upstream of the MTase transcription initiation site for the presence of conserved sequence patterns, especially those found in eukaryotic promoters. This analysis highlighted the presence of putative eukaryotic promoter elements: TATA boxes were recovered in almost all the genomes analysed, whereas one mismatch CAAT boxes were found in 23% of the genomes and one mismatch GC boxes in 23% of them (Figure 4). The three eukaryotic transcription elements are never found simultaneously on the same promoter, but TATA boxes are often associated with CAAT or with GC elements. The genomic region upstream the MTase gene in wDi (part of the RNA-Seq studies included in this work) resulted to contain a perfect TATA box but it was not recognised as a eukaryotic-like promoter by the Neural Network Promoter Prediction tool (Figure 4). This highlights that the regulation of the expression of the MTase gene remains to be investigated more deeply in the future. Interestingly, in eukaryotes, genes are classified based on the presence of the TATA box at promoter regions, in TATA-containing- and TATA-less genes [39,40]. While the latter are mostly highly expressed genes with survival-related housekeeping functions, TATA-containing genes are a minority, representing 20% of genes in *Saccharomyces cerevisiae* and about 25% of genes in humans [39]. They tend to be stress-responsive genes, whose expression must rapidly and variably be tuned in response to specific environmental conditions and changing physiological needs [41,42] . Thus, the host could overexpress TATA box-associated transcription factors, and change whole Wolbachia gene expression, under environmental stress conditions.

Our hypothesis is that host organisms may control whole Wolbachia gene expression by modulating the transcription of its MTase gene, via a eukaryotic-like promoter located upstream of that gene. The expressed MTase enzyme would then unspecifically methylate cytosines and adenines along the Wolbachia genome. Then, methylation would affect the transcription of all Wolbachia genes, enhancing the genes with higher cytosines / adenine ratio (Figure 6). Using the TATA box as the main regulatory element, this mechanism could act as a Wolbachia-dependent host response to stressful conditions. This idea is also

sustained by the annotation of the genes with high or low G/T ratio (possibly more affected by this mechanisms, as explained above): the genes with low G/T ratio are mainly involved in basal metabolism pathways (such as replication and translation) while the high G/T ratio genes are mostly part of pathways involved in stress response (such as the production of energy and ion/molecule transport). Interestingly, among the low or high G/T ratio genes there are, respectively, *fts*Z and *fts*W, both involved in the bacterium cell replication. This suggests that the host could modulate the bacterium replication: a possible key for a successful symbiotic relationship. In this picture, Wolbachia could have a stress-related mitochondrion-like role for the host, as previously proposed by Darby and colleagues (2012).

While this mechanism may not lead to highly precise regulation of gene expression, it could be a rough yet functional mechanism, sufficient for the establishment of a successful intracellular symbiosis. In the future, it would be interesting to investigate whether similar mechanisms are present in other endosymbiont or intracellular organisms.

Despite the results presented in this work are far from being definitive, we hope to have opened a new way in understanding the symbiotic relationship between Wolbachia and its hosts. In our opinion, this is a magnificent example of how evolution works, recycling metabolic pieces for similar aims, sometimes reaching surprisingly stable *equilibria*.

Captions

Figure 1. Methylation patterns of the DNA methyltransferase gene of Wolbachia endosymbiont of *Onchocerca ochengi* (wOo)

Sequence logo of the sequences flanking (from -20 to +20 positions) the methylated positions found in the genome of Stellar *Escherichia coli* dam-/dcm- strain expressing the DNA methyltransferase gene of Wolbachia endosymbiont of *Onchocerca ochengi* (wOo). The height of the letters depends on the conservation of the position among the sequences. a) the sequence logo of the m4C methylations on cytosines; b) the m6A methylation of adenine.

Figure 2. Correlation between the expression of genes and their %G / %T ratio

For each of the 55 samples included in the study, the correlation between the expression of the genes and the relative %G/%T ("G/T ratio") was investigated. The scatter plots of G/T ratio vs gene expression of three exemplificative samples selected from the 55 samples are reported: a) the correlation plot with highest slope (Hs), b) the correlation plot with the medium slope (Ms) and c) the correlation plot with the lowest slope (Ls). In each of these three plots, colour is green whether the correlation is significant (p value < 0.05) and grey if it is not, the regression line is coloured in dark grey and the statistics are reported on the chart. In figure d, the correlation plot between the slopes obtained for each of the 55 samples, and the relative expression of the DNA methyltransferase gene is reported. Points relative to significant correlations (p value < 0.05) are in green while non significant in grey. The regression line is in green and the statistics are reported on the chart.

Figure 3. Correlation between the expression of genes and their Codon Bias Index (CBI)

For each of the 55 samples included in the study, the correlation between the expression of the genes and the relative Codon Bias Index (CBI) was investigated. The scatter plots of CBI

vs gene expression of three exemplificative samples selected from the 55 samples are reported: a) the correlation plot with highest slope (Hs), b) the correlation plot with the medium slope (Ms) and c) the correlation plot with the lowest slope (Ls). In each of these three plots, colour is green whether the correlation is significant (p value < 0.05) and grey if it is not, the regression line is coloured in dark grey and the statistics are reported on the chart. In figure d, the correlation plot between the slopes obtained for each of the 55 samples, and the relative expression of the DNA methyltransferase gene is reported. Points relative to significant correlations (p value < 0.05) are in green while non significant in grey. The regression line is in green and the statistics are reported on the chart.

Figure 4. TATA boxes in the 200 bp regions upstream the DNA methyltransferase gene

The 200 bp regions upstream of the DNA methyltransferase gene (MTase) on 111 representative Wolbachia strains are shown. On the left, the Maximum Likelihood (ML) phylogenetic analysis obtained from the concatenate of the single copy core genes. On the right, the position of perfect boxes are shown with different colours: TATA boxes in red, CAAT boxes in green, GC boxes in orange and Pribnow boxes in blue. One mismatch boxes are also shown using the same colours but in transparency. The names of the upstream regions found to contain putative eukaryotic promoters by the Neural Network Promoter Prediction tool are bolded, and the names of the wDi, wOo, wBm and wMel strains used in the genomic and statistical analyses of the work (see material and methods) are in blue. Lastly, one sequence (GCF_018454445.1, Wolbachia endosymbiont of Rhagoletis cerasi - wCer5) is composed only by gap because the MTase gene was placed on the extreme of the contig and it was not possible to retrieve the upstream region.

Figure 5. COG annotated genes with highest or lowest G/T ratio

The heatmap shows the pattern of the genes with high or low G/T ratio among the 112 Wolbachia strains included in the study. On the left, the Maximum Likelihood phylogenetic

tree obtained from concatenate of the single copy core genes. In the middle, heatmap reports the pattern of the 10 lower G/T ratio COG annotated genes: rows refer to strains (matching with the phylogenetic tree on the left) and columns to genes (COG annotation is reported). Cell colours report the G/T ratio value following the legend on lower left. Asterisks are reported for genes among those in the 10 lower G/T ratio for the relative strain. More on the right, a heatmap reports the 10 higher G/T ratio COG annotated genes. Finally, on the right, Wolbachia strains hosts and supergroups are shown.

Figure 6. Graphical representation of the mechanism proposed for the regulation of the Wolbachia gene expression

a) The grey circle represents the Wolbachia genome. A putative host transcription factor (orange triangle) binds the eukaryotic-like promoter (in yellow on the Wolbachia genome) upstream of the Wolbachia SAM-dependent DNA methyltransferase (MTase) gene (in red on the Wolbachia genome) leading to the expression of the gene. The MTase enzyme is represented in red on the right. b) The MTase enzyme methylates the adenines and cytosines of the Wolbachia DNA and methyl groups are represented with dots coloured as the bases. c) RNA polymerases are represented by pink icons from which mRNA molecules originate. In highly expressed genes more RNA polymerases bind the gene simultaneously. The greater is the ratio between the number of cytosines or adenines in a gene (corresponding to guanines and thymines in the corresponding Open Reading Frames and mRNAs), the more that gene is expressed.

Funding

The study was supported by the National Institute of Virology and Bacteriology (Program EXCELES, ID project no. LX22NPO5103) funded by the European Union–Next Generation EU.

References

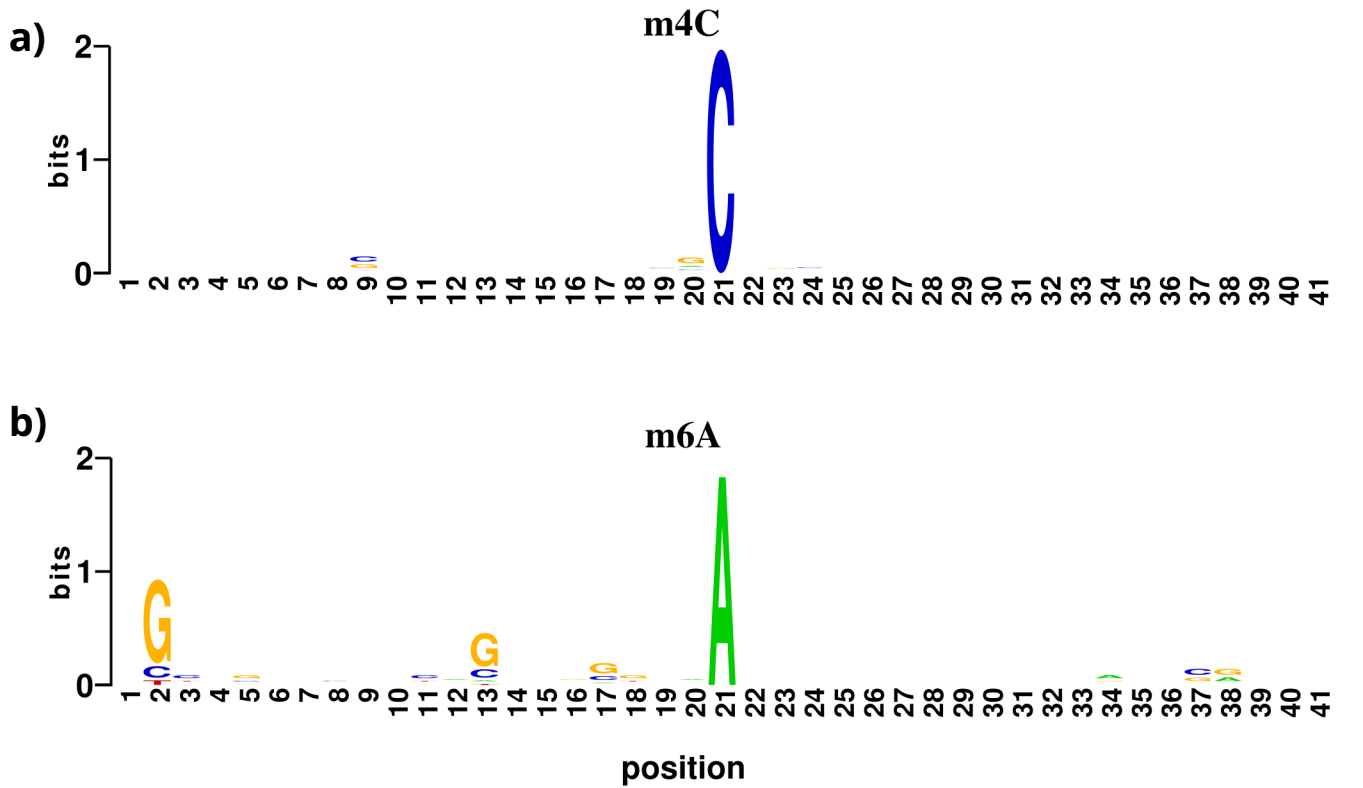
1. Werren, J. H., Baldo, L. & Clark, M. E. Wolbachia: master manipulators of invertebrate biology. *Nat. Rev. Microbiol.* **6**, 741–751 (2008).
2. Zug, R. & Hammerstein, P. Still a host of hosts for Wolbachia: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. *PLoS One* **7**, e38544 (2012).
3. Porter, J. & Sullivan, W. The cellular lives of Wolbachia. *Nat. Rev. Microbiol.* **21**, 750–766 (2023).
4. Werren, J. H. Biology of Wolbachia. *Annu. Rev. Entomol.* **42**, 587–609 (1997).
5. Hedges, L. M., Brownlie, J. C., O'Neill, S. L. & Johnson, K. N. Wolbachia and virus protection in insects. *Science* **322**, 702 (2008).
6. Teixeira, L., Ferreira, A. & Ashburner, M. The bacterial symbiont Wolbachia induces resistance to RNA viral infections in Drosophila melanogaster. *PLoS Biol.* **6**, e2 (2008).
7. Newton, I. L. G. & Rice, D. W. The Jekyll and Hyde Symbiont: Could Be a Nutritional Mutualist? *J. Bacteriol.* **202**, (2020).
8. Bandi, C. *et al.* Effects of tetracycline on the filarial worms Brugia pahangi and Dirofilaria immitis and their bacterial endosymbionts Wolbachia. *Int. J. Parasitol.* **29**, 357–364 (1999).
9. Scholz, M. *et al.* Large scale genome reconstructions illuminate Wolbachia evolution. *Nat. Commun.* **11**, 5235 (2020).
10. Tolley, S. J. A., Nonacs, P. & Sapountzis, P. Horizontal Transmission Events in Ants: What Do We Know and What Can We Learn? *Front. Microbiol.* **10**, 296 (2019).
11. Olson, R. D. *et al.* Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* **51**, D678–D689 (2023).

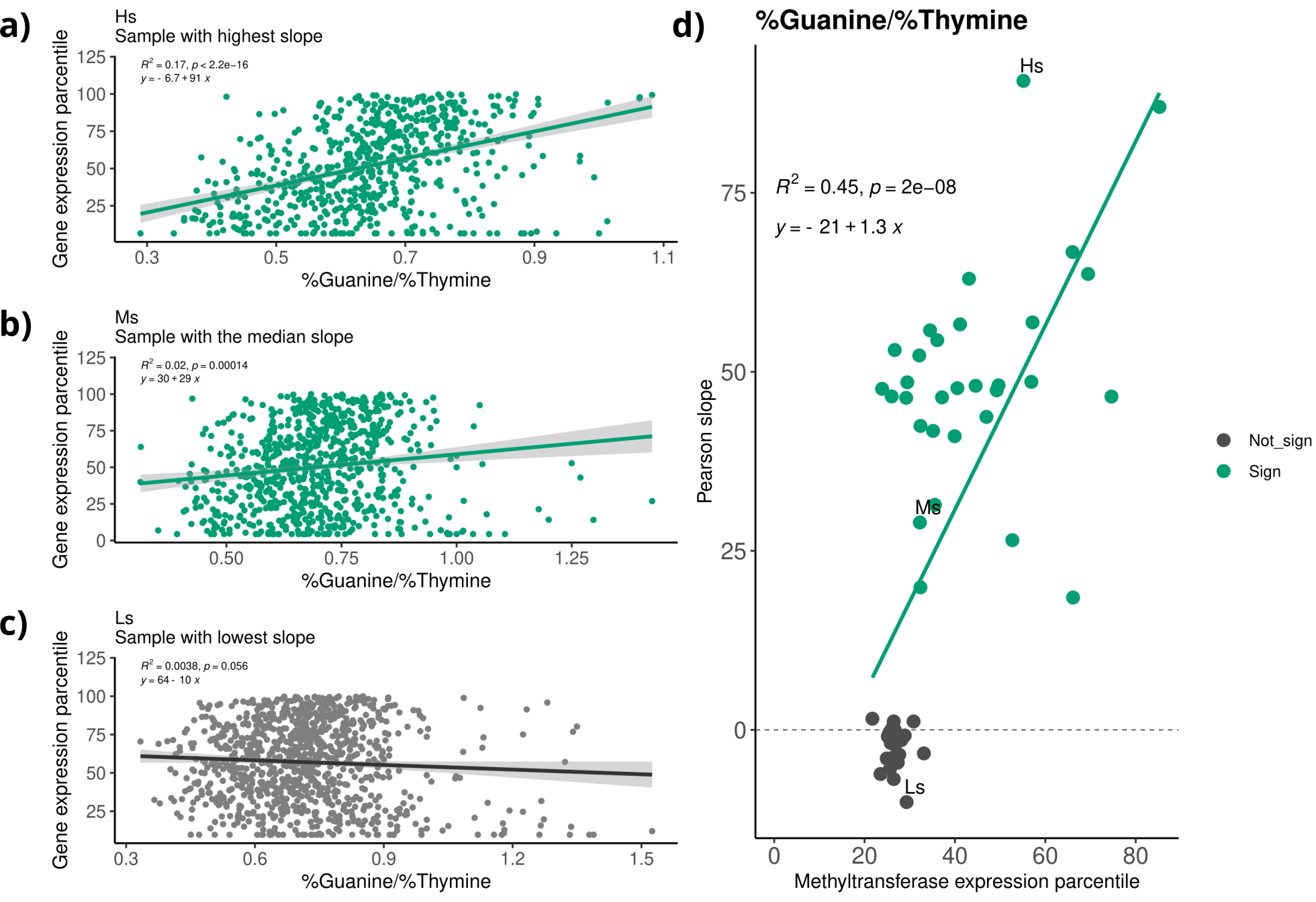
12. Manzano-Marín, A. & Latorre, A. Snapshots of a shrinking partner: Genome reduction in Serratia symbiotica. *Sci. Rep.* **6**, 32590 (2016).
13. Comandatore, F. *et al.* Supergroup C Wolbachia, mutualist symbionts of filarial nematodes, have a distinct genome structure. *Open Biol.* **5**, 150099 (2015).
14. Mahmood, S., Nováková, E., Martinů, J., Sychra, O. & Hypša, V. Supergroup F Wolbachia with extremely reduced genome: transition to obligate insect symbionts. *Microbiome* **11**, 22 (2023).
15. Darby, A. C. *et al.* Analysis of gene expression from the Wolbachia genome of a filarial nematode supports both metabolic and defensive roles within the symbiosis. *Genome Res.* **22**, 2467–2477 (2012).
16. Luck, A. N. *et al.* Concurrent transcriptional profiling of Dirofilaria immitis and its Wolbachia endosymbiont throughout the nematode life cycle reveals coordinated gene expression. *BMC Genomics* **15**, 1041 (2014).
17. Darby, A. C. *et al.* Integrated transcriptomic and proteomic analysis of the global response of Wolbachia to doxycycline-induced stress. *ISME J.* **8**, 925–937 (2014).
18. Luck, A. N. *et al.* Tissue-specific transcriptomics and proteomics of a filarial nematode and its Wolbachia endosymbiont. *BMC Genomics* **16**, 920 (2015).
19. Gutzwiller, F. *et al.* Dynamics of Wolbachia pipientis Gene Expression Across the Drosophila melanogaster Life Cycle. *G3* **5**, 2843–2856 (2015).
20. Grote, A. *et al.* Defining Brugia malayi and Wolbachia symbiosis by stage-specific dual RNA-seq. *PLoS Negl. Trop. Dis.* **11**, e0005357 (2017).
21. Chung, M. *et al.* Drug Repurposing of Bromodomain Inhibitors as Potential Novel Therapeutic Leads for Lymphatic Filariasis Guided by Multispecies Transcriptomics. *mSystems* **4**, (2019).
22. Papaleo, S. *et al.* The red thread between methylation and mutation in bacterial antibiotic resistance: How third-generation sequencing can help to unravel this relationship. *Front. Microbiol.* **13**, 957901 (2022).
23. Sánchez-Romero, M. A. & Casadesús, J. The bacterial epigenome. *Nat. Rev. Microbiol.*

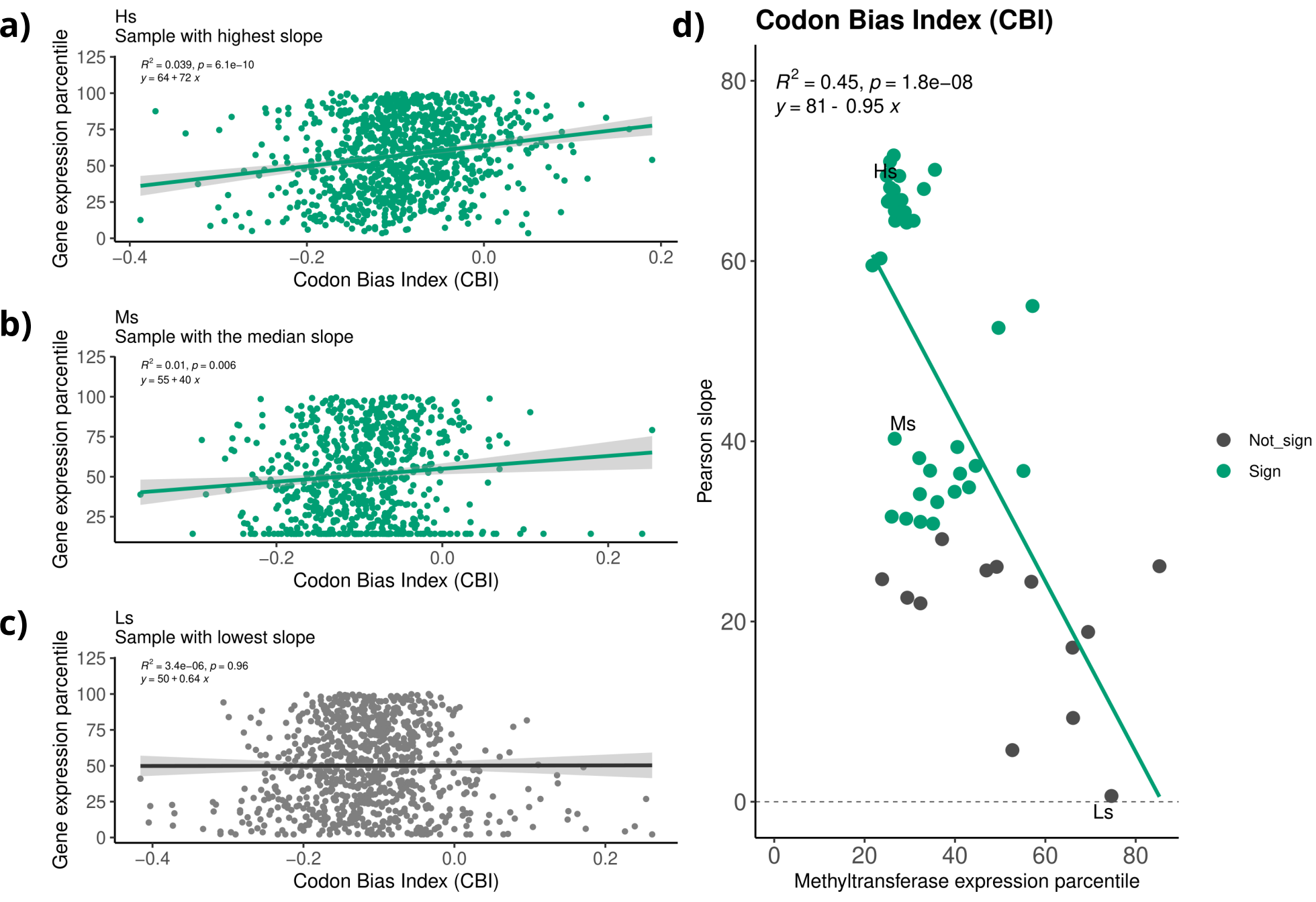
18, 7–20 (2020).

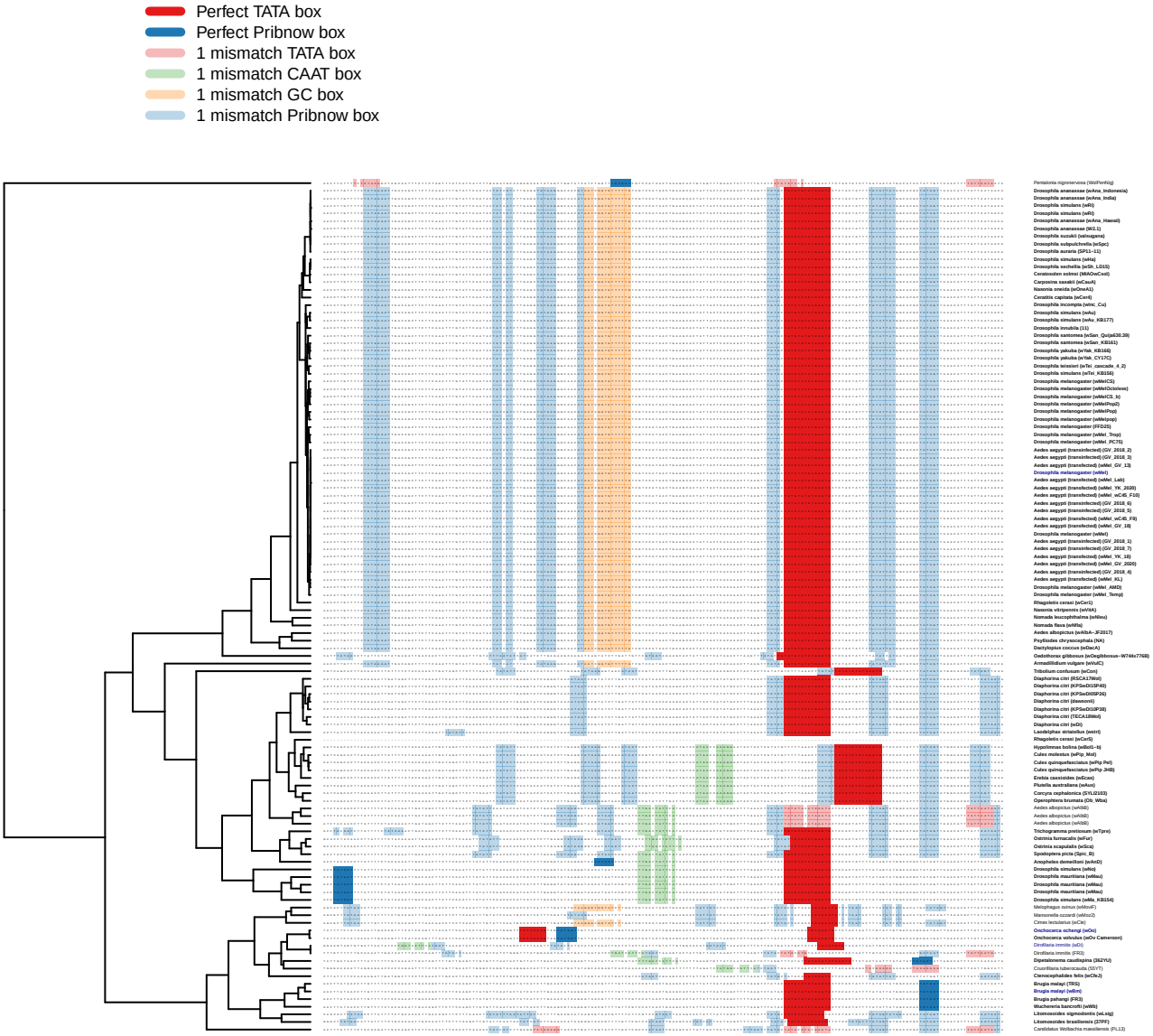
24. Chung, M. *et al.* A Meta-Analysis of Transcriptomics Reveals a Stage-Specific Transcriptional Response Shared Across Different Hosts. *G3* **10**, 3243–3260 (2020).
25. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
26. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
27. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
28. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006).
29. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
30. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
31. Darriba, D. *et al.* ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.* **37**, 291–294 (2020).
32. Revell, L. J. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* **12**, e16505 (2024).
33. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**, 563–578 (1990).
34. Comandatore, F. *et al.* Phylogenomics and analysis of shared genes suggest a single transition to mutualism in Wolbachia of nematodes. *Genome Biol. Evol.* **5**, 1668–1674 (2013).
35. Dale, C. & Moran, N. A. Molecular Interactions between Bacterial Symbionts and Their Hosts. *Cell* **126**, 453–465 (2006).

36. Gao, Q. *et al.* Bacterial DNA methyltransferase: A key to the epigenetic world with lessons learned from proteobacteria. *Front. Microbiol.* **14**, 1129437 (2023).
37. Jensen, T. Ø. *et al.* Genome-wide systematic identification of methyltransferase recognition and modification patterns. *Nat. Commun.* **10**, 3311 (2019).
38. Quax, T. E. F., Claassens, N. J., Söll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell* **59**, 149–161 (2015).
39. Han, H. W., Bae, S. H., Jung, Y. H., Kim, J. H. & Moon, J. Genome-wide characterization of the relationship between essential and TATA-containing genes. *FEBS Lett.* **587**, 444–451 (2013).
40. López-Maury, L., Marguerat, S. & Bähler, J. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat. Rev. Genet.* **9**, 583–593 (2008).
41. Basehoar, A. D., Zanton, S. J. & Pugh, B. F. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**, 699–709 (2004).
42. Bae, S.-H., Han, H. W. & Moon, J. Functional analysis of the molecular interactions of TATA box-containing genes and essential genes. *PLoS One* **10**, e0120848 (2015).


a)
m4C
bits
b)
m6A
bits
position

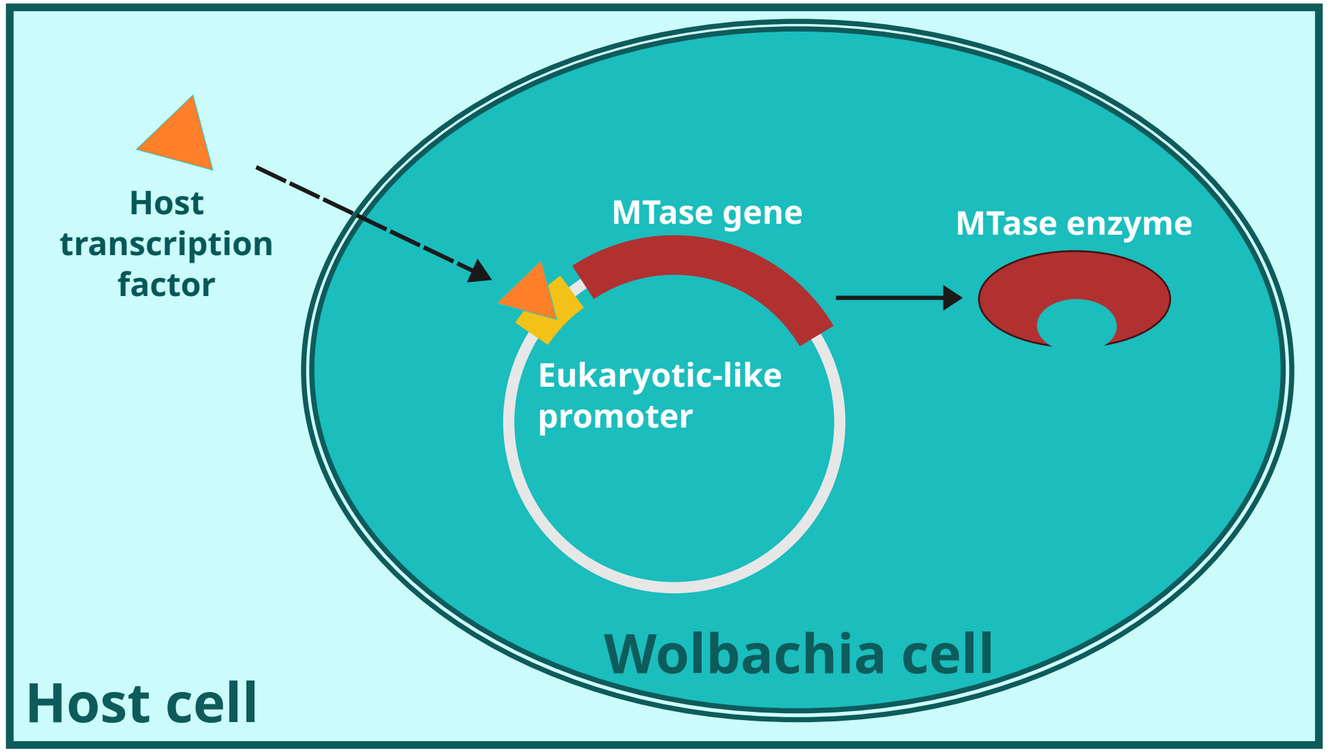

a)
Hs
Sample with highest slope
$R^2 = 0.17, p < 2.2e-16$
$y = -6.7 + 91x$
Gene expression parcentile
%Guanine/%Thymine
b)
Ms
Sample with the median slope
$R^2 = 0.02, p = 0.00014$
$y = 30 + 29x$
Gene expression parcentile
%Guanine/%Thymine
c)
Ls
Sample with lowest slope
$R^2 = 0.0038, p = 0.056$
$y = 64 - 10x$
Gene expression parcentile
%Guanine/%Thymine
d)
%Guanine/%Thymine
Hs
Ms
Ls
$R^2 = 0.45, p = 2e-08$
$y = -21 + 1.3x$
Pearson slope
Methyltransferase expression parcentile
Not_sign
Sign


a)
Hs
Sample with highest slope
R^2 = 0.039, p = 6.1e−10
y = 64 + 72 x
Gene expression parcentile
Codon Bias Index (CBI)
b)
Ms
Sample with the median slope
R^2 = 0.01, p = 0.006
y = 55 + 40 x
Gene expression parcentile
Codon Bias Index (CBI)
c)
Ls
Sample with lowest slope
R^2 = 3.4e−06, p = 0.96
y = 50 + 0.64 x
Gene expression parcentile
Codon Bias Index (CBI)
d)
Codon Bias Index (CBI)
R^2 = 0.45, p = 1.8e−08
y = 81 - 0.95 x
Hs
Ms
Ls
Pearson slope
Methyltransferase expression parcentile
Not_sign
Sign

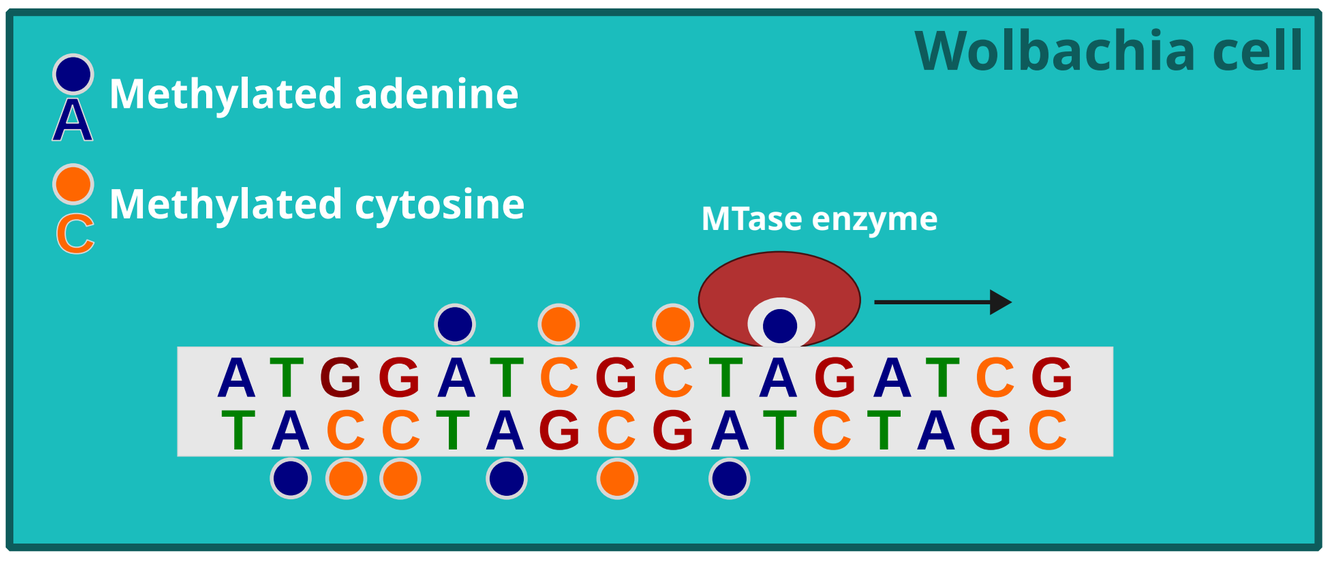
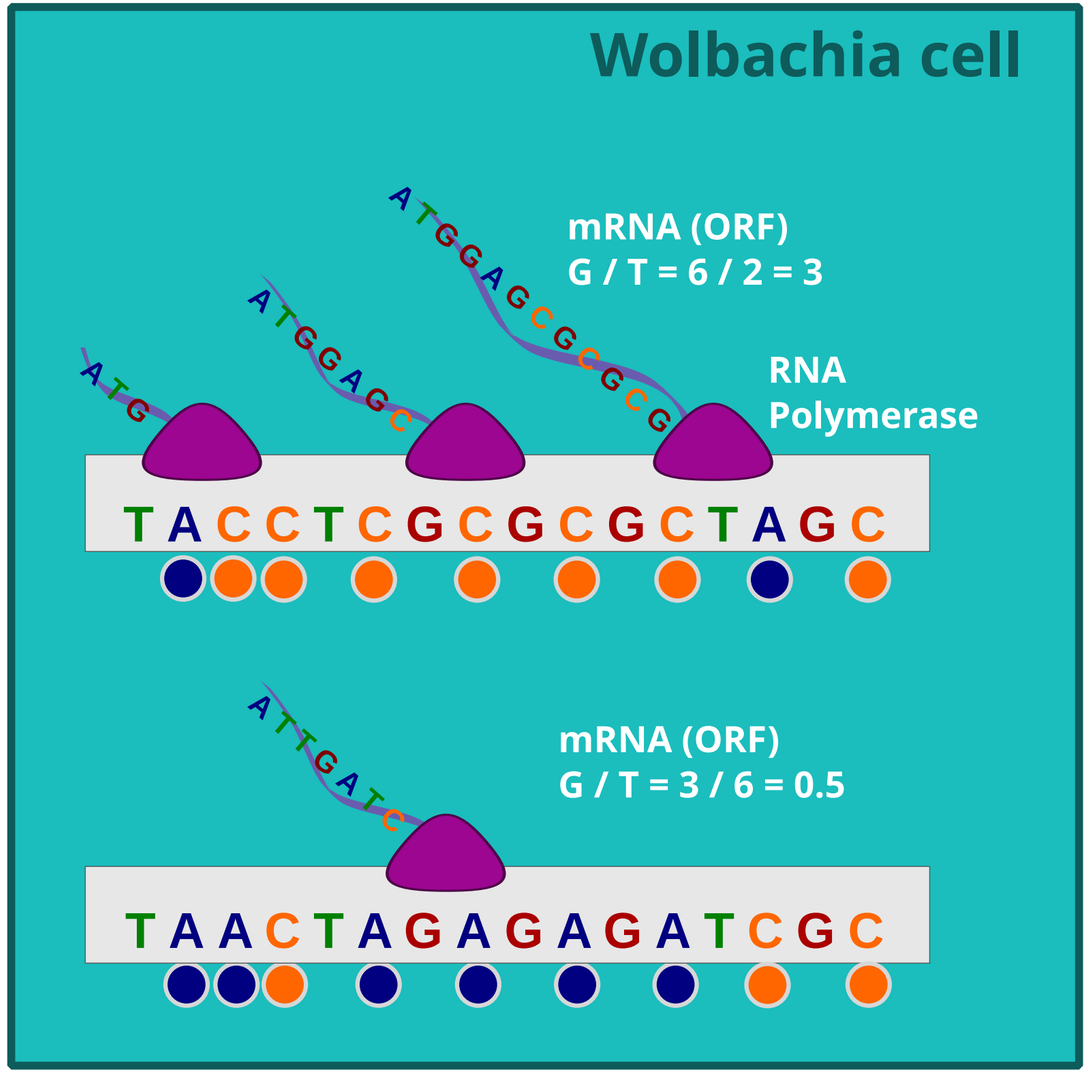

Perfect TATA box
Perfect Pribnow box
1 mismatch TATA box
1 mismatch CAAT box
1 mismatch GC box
1 mismatch Pribnow box

Lowest G/T
Highest G/T
M
A
B
F
C
J
I
D
T
G/T ratio
Count
0
1000
2500
1
2
3
Value

a) Host induces MTase expression



b) MTase enzyme methylates adenines and cytosines



c) Genes with ORFs with high G/T are more expressed