The Transcriptional landscape of *Streptococcus pneumoniae* reveals a complex operon architecture and abundant riboregulation critical for growth and virulence

Indu Warrier†, Nikhil Ram-Mohan†, Zeyu Zhu, Ariana Hazery, Michelle M Meyer*, Tim van Opijnen*

Biology Department, Boston College, Chestnut Hill, MA, USA

† Equal contribution
* Corresponding author

E-mail Addresses:

MMM: m.meyer@bc.edu
TvO: vanopijn@bc.edu

Keywords:
RNA-Seq
term-seq
5'end-seq
Riboregulation
sRNA
transcriptional landscape
operon architecture
Streptococcus pneumoniae

Abstract

Efficient and highly organized transcription initiation and termination is fundamental to an organism's ability to survive, proliferate, and quickly respond to its environment. Over the last decade, our simplistic outlook of bacterial transcriptional regulation and architecture has evolved to include stimulus-responsive regulation by untranslated RNA and the formation of alternate transcriptional units. In this study, we map the transcriptional landscape of the bacterial pathogen *Streptococcus pneumoniae* by applying a combination of high-throughput RNA-sequencing techniques. Our study reveals a complex transcriptome wherein environment-respondent alternate transcriptional units are observed within operons stemming from internal transcription start sites (TSS) and transcription terminators (TTS) suggesting that more fine-tuning of regulation occurs than previously thought. Additionally, we identify many putative *cis*-regulatory RNA elements and riboswitches within 5'-untranslated regions (5'-UTR) of genes. By integrating TSSs and TTSs with independently collected RNA-Seq datasets from a variety of conditions, we establish the response of these regulators to changes in growth conditions and validate several of them. Furthermore, to determine the importance of ribo-regulation by 5'-UTR elements for *in vivo* virulence, we show that the pyrR regulatory element is essential for survival, successful colonization and infection in mice suggesting that such RNA elements are potential drug targets. Importantly, we show that our approach of combining high-throughput sequencing with *in vivo* experiments can reconstruct a global understanding of regulation, but also pave the way for discovery of compounds that target (ribo-)regulators to mitigate virulence and antibiotic resistance.

Introduction

The transcriptional architecture of bacterial genomes is far more complex than originally proposed. The classical model of an operon describes a group of genes under the control of a regulatory protein where transcription results in a polycistronic mRNA with a single transcription start site (TSS) and a single transcription terminator site (TTS) [1]. However, many individual examples have established that the same operon may encode alternative transcriptional units under varying environmental conditions [2,3]. Furthermore, advancements in sequencing technology that enable highly accurate mapping of TSSs and TTSs on a genome-wide level have demonstrated that the number of TSSs and TTSs can significantly exceed the number of operons [4]. Thus it seems likely that the bacterial transcriptional landscape, or the genome-wide map of all possible transcriptional units, is shaped by an operon architecture that encodes many TSSs and TTSs within single operons, thus significantly increasing complexity with the objective of enabling diverse transcriptional outcomes [5,6].

To achieve a complex landscape of alternative transcriptional units, transcriptional regulation occurs on multiple levels. In addition to the many protein activators and repressors that control transcription initiation, there are also many non-coding RNAs (ncRNAs), including both small ncRNAs (sRNAs) and highly structured portions of mRNAs that play essential roles as regulatory elements controlling metabolism, stress-responses, and virulence [7,8]. *Trans* – acting small RNAs (sRNAs) allow selective degradation or translation of specific mRNAs [9] and *cis* – acting mRNA structures, such as riboswitches, interact with small molecules including, metal ions, and protein ligands to affect expression of downstream genes by regulating transcription attenuation or translation inhibition [10]. RNA regulation has been shown to play a key role in

77 shaping the transcriptional landscape of a wide range of pathogenic bacteria including
78 *Staphylococcus aureus*, *Listeria monocytogenes*, *Helicobacter pylori*, and several strains of
79 *Streptococci* [6,11-19]. Several RNA regulators have been validated and associated with
80 pathogenicity and virulence [20,21], and could be used as highly specific druggable targets
81 [22,23], however, only a select set of regulators have been targeted to date [24,25].
82
83 *Streptococcus pneumoniae* is a major causative agent of otitis media, meningitis, pneumonia, and
84 bacteremia. It causes 1.2 million cases of drug-resistant infections in the US annually and results
85 in ~1 million deaths per year worldwide [26-28]. While high-resolution transcriptional mapping
86 data are available for other *Streptococcus* species, these studies have shown limited experimental
87 validation [17], or have focused primarily on the role of sRNAs in virulence [29]. Additionally,
88 previous studies of the *S. pneumoniae* transcriptome have demonstrated the presence of ncRNA
89 regulators and assessed their roles in infection and competence, however, these studies also
90 largely focused on sRNAs [13,15,30]. Thus, while potentially incredibly valuable, a high-
91 resolution validated map of the genome-wide transcriptional landscape for *S. pneumoniae* is not
92 available.
93
94 Here a comprehensive characterization of the *S. pneumoniae* TIGR4 transcriptional landscape is
95 created using RNA-Seq [31], 5' end-Seq [19], and term-seq (3'-end sequencing) [32]. We obtain
96 a global transcript coverage map, identifying all TSSs, and all TTSs, which highlights a highly
97 complex *S. pneumoniae* transcriptional landscape including many operons with multiple TSSs
98 and TTSs. Furthermore, we demonstrate how TSS and TTS mapping under one set of conditions
99 can be leveraged to analyze independently obtained RNA-Seq data collected under a variety of

conditions, and we experimentally validate this approach with several cis-acting RNA regulators. Finally, we demonstrate that the functionality provided by the RNA *cis* – regulator *pyrR* is critical for *S. pneumoniae in vivo* in a mouse infection model. Importantly, our work demonstrates how a variety of high-throughput sequencing efforts can be combined to map out a comprehensive transcriptional landscape for a bacterial pathogen as well as identify potentially druggable ncRNA targets.

Results:

Streptococcus pneumoniae has a complex transcriptional landscape.

To characterize the transcriptional landscape of *Streptococcus pneumoniae* TIGR4 (T4), we first determined transcript boundaries by mapping transcription start (TSSs) and termination sites (TTSs) from 5' and 3' end sequencing reads obtained from exponential growth (Fig 1A and Fig 1B). For the 2341 annotated genes in T4, a total of 1597 TSSs and 1330 TTSs were identified, as well as 236 antisense terminators suggesting extensive antisense regulation (Fig 2, S1 and S2 Tables). RNA-Seq based clustering of genes with Rockhopper [33] detected 773 single gene operons and grouped 1512 genes into 474 multi-gene operons (S3 Table, and Fig 3). We classified operons into five categories based on the number of internal TSSs and TTSs (Fig 2). The majority of *S. pneumoniae* genes are independent transcriptional units with a single TSS and TTS (simple operons) (Fig 3A). Traditional operons with multiple genes and a single TSS and TTS make up 5% of the operons, while multiTSS, and multiTTS operons make up 4% and 3% of transcriptional units respectively. However, complex operons (most of which consist of two genes) with multiple TSSs and TTSs are the second largest category comprising 26% of all operons (Fig 3A and Fig 3B). Most complex operons are defined by a secondary internal TSS

and TTS, however there are several significantly more complex examples where the operon contains multiple TSSs and TTSs (Fig 3B), indicating a highly intricate system of possible transcripts.

Since our data revealed many operons with complex structure, we sought to corroborate specific examples using additional data sources. One complex operon we identified, which is also present in existing databases of operon structure [5,34], consists of 9 genes (SP1018-SP1026) with six internal TSSs and eight TTSs (Fig 4A). In addition, this operon displays unequal and complex gene expression patterns when independently collected RNA-Seq data from diverse media conditions is mapped to the transcript. In poor growth medium (MCDM) the operon can be split into two parts based on expression, where the last five genes in the operon (SP1022-1026) are expressed significantly higher than the first four genes (SP1018-1021), while in rich medium (SDMM) the read depth across the operon is similar. This observation corroborates the role of the internal regulatory mechanisms for maintaining differences in gene expression between different growth conditions.

A second validation of our data and analysis approach derives largely from existing low-throughput experiments. The mal regulon is a multiple operon system under the control of a single protein, malR (SP2112), which downregulates regulon expression at the malM (SP2107) promoter. Our data shows that the mal regulon includes operons belonging to three different categories, a traditional operon (malAR/SP2111-2112), a multiTTS operon (malMP/SP2016-2107) and a complex operon (malXCD/SP2108-2110). From the RNA-Seq coverage maps it is clear that the three operons can be differentially controlled and expressed in rich vs poor media

(Fig 4B). Furthermore, the TSS and TTS identified by our analysis reveal features that have been previously described in lower throughput assays [24]. Thus, although our data may highlight many examples of complex transcriptional architecture, these examples are verifiable through the incorporation of additional RNA-Seq data, and where applicable are consistent with low-throughput studies done in the past.

Genome-wide identification and pan-genome wide conservation of regulatory RNAs.

To identify RNA regulators that act through premature transcription termination, we compiled 3' end sequencing reads upstream of translational start sites, allowing a minimum 5'-untranslated region (UTR) length of 70 bases. We detected 565 such early TTS sites that represent putative regulatory elements (represented in black (TSS) and orange (early TTS) in the inner band of Fig 2). By screening these regulatory elements against 380 published *S. pneumoniae* strains [35], covering a large part of the pan-genome, we found that 20 candidates (~4%) were conserved across all genomes, 171 (45%) were identified in at least 350 genomes, 68 (~12%) candidates were identified in fewer than half of the genomes (S1A Fig), while a single candidate, identified upstream of a transposase, was found only in T4. Interestingly, 415 (73%) candidates were found as single copies within a genome, while the others had varying copy numbers ranging from 2 to 29. Evolutionary distance of each candidate cluster was estimated using MEGA-CC [36], which reveals that each cluster is made of highly similar, if not identical, sequences (S1B Fig).

The candidate RNAs were compared to previously identified *S. pneumoniae* small RNAs and to homologs of characterized structured RNA families (including a variety of riboswitches and other *cis*-regulatory structured RNAs). A total of 111 candidates overlap with previously

identified *S. pneumoniae* small RNAs (sRNAs), 86 out of the 88 intergenic sRNAs described in [13] and 32, with an additional 24 within 150 nucleotides, out of 89 described by [15] (S5 Table). To identify homologs of characterized RNA families in T4 we used Infernal (an RNA specific homology search tool, [37]) to search the genome, which detected 51 of the 565 candidates, and 3 out of the 6 regulatory elements experimentally validated in this work (S4 Table), indicating the existence of many novel regulatory elements in T4. However, the coordinates identified by Infernal do not always match with those identified by our methodology (S4 Table). For example, the 23S-methyl ncRNA identified by Infernal is found in T4 between coordinates 466473 and 466568, and the regulatory element experimentally identified is found between coordinates 466469 and 466579. All the Infernal identified ncRNAs overlapping with the candidates identified here are listed in S4 Table. On the other hand, in certain cases like the L1 regulator, the coordinates do not overlap at all. This could be due to the existence of a condition specific secondary TSSs that were not picked up in our growth conditions. Despite these exceptions, the majority of the Infernal identified ncRNAs families show complete overlap with the candidates.

Leveraging RNA-Seq data collected under various conditions enables identification and validation of environment-responsive RNA regulators.

We reasoned that since we mapped RNA-regulators by means of 5'end sequencing, we would be able to associate these regulators with specific growth conditions using environment dependent RNA-Seq data. To confirm the biological relevance of an RNA regulator and associate it with a specific condition one would expect to see a change in the 5' UTR coverage relative to the accompanying gene. For instance, if a regulator forms an early terminator the RNA-Seq coverage in the 5'UTR is relatively high, while the coverage in the controlled gene would be

much lower. Alternatively, if the environment relieves the formation of the early terminator the coverage across the 5'UTR and gene would become less skewed. To determine the applicability of this assumption we leveraged independently collected RNA-Seq data sampled under different nutrient conditions including, rich and poor media, and nutrient depletion conditions where a single nutrient was removed from the environment. RNA-Seq data were mapped to each putative regulatory region and coverage was calculated and averaged across the length of the 5' UTR regulatory element and the downstream gene. From our list of candidate 5'-UTR regulatory elements, 128 showed more than two fold change in read-through between rich and poor media, with the majority showing an increase in read-through in poor media. Five regulators from this list were validated by qRT-PCR (Fig 5A-B, S2 Fig), confirming that RNA-Seq data can indeed be used to identify conditions to which an RNA regulator responds.

Importantly, two of the validated regulators are known as the thiamine pyrophosphate (TPP) and flavin mononucleotide (FMN) riboswitches. In many bacteria the TPP riboswitch binds thiamine pyrophosphate and regulates thiamine biosynthesis and transport [38]. Similarly, the FMN riboswitch regulates biosynthesis and transport of riboflavin by binding to FMN [39]. While we validated that these riboswitches respond to poor media by increasing expression of their respective genes (Fig 5A-B) we suspected that this was due to depletion of each specific ligand in the poor media. Indeed, when poor media is supplemented either with thiamine or riboflavin, expression of the TPP or FMN controlled gene (SP0716 and SP0178 respectively) decreases by more than 3-fold, suggesting that the observed differences between rich and poor media can be attributed to the activity of these riboswitches.

In an attempt to validate the feasibility of directly associating RNA-regulators with a highly specific change in the environment we performed RNA-Seq in the presence and absence of uracil. One specific regulatory element that is sensitive to uracil is the pyrR RNA element, which in many bacteria regulates *de novo* pyrimidine nucleotide biosynthesis through a transcription attenuation mechanism mediated by the PyrR regulatory protein [40,41]. In the presence of the co-regulator UMP, PyrR binds to the 5' UTR of the *pyr* mRNA transcript (the pyrR RNA element) and disrupts the anti-terminator stem-loop thereby promoting the formation of a factor-independent transcription terminator resulting in reduced expression of downstream genes [40] (Fig 6A). In contrast, the co-regulator 5-phosphoribosyl-1-pyrophosphate (PRPP) antagonizes the action of UMP on termination by binding to the PyrR protein when UMP concentration is low [42] (Fig 6A). For *S. pneumoniae*, our data confirms that pyrR RNA elements are present in the 5' UTR of two *pyr* operons (SP1278-1276; SP0701-0702), and the uracil transporter (SP1286). Furthermore, in response to the absence of uracil the coverage across the two genes directly adjacent to the regulators (SP1278 and SP0701) and over the entire two operons increases drastically (Fig 6C), demonstrating that the regulatory elements effectively turn the genes/operons on, which we confirmed by qRT-PCR (Fig 6D, S3 Fig). Thus, while term-seq can be used to map novel regulatory RNA candidates on a genome-wide scale, RNA-Seq data can be leveraged, even in retrospect, to identify environmental conditions the regulator responds to.

The pyr operon is regulated through the secondary structure of the 5' RNA leader-region, is essential for in vitro growth and in vivo virulence and can be directly manipulated.

To further investigate the importance of the pyrR regulatory RNA element in growth, three different mutants were constructed that variably affect the 5' RNA secondary structure (Fig 7A):

1) mutation M1 interferes with the binding of PyrR to the *pyr* mRNA; 2) mutation M2 renders the regulatory element in an "always on" state by destabilizing the rho-independent terminator stem-loop structure that is formed in the presence of UMP; 3) M3 locks the terminator and creates an "always off" state (Fig 7A). Wild type and mutant strains were cultured in the presence or absence of uracil and the effect of the mutations on expression of SP1278 were assessed with qRT-PCR (Fig 7B). As expected, expression in the wild type decreased (9.5-fold) in the presence of uracil confirming the repressive effect of exogenous pyrimidine (Fig 7B) [43]. M1, which should be insensitive to the presence of PyrR and its co-regulator UMP (Fig 7A) is indeed unresponsive to the presence of uracil (Fig 7B). M2 triggers constitutive expression of the *pyr* operon (Fig 7B) and M3 has a ~5-fold reduction in expression compared to the wild type regardless of the presence of uracil (Fig 7B).

Previously we showed that the pyrimidine synthesis pathway in *S. pneumoniae* is partially regulated by a two-component system (SP2192-2193) and that genes in this pathway are important for growth [44]. To determine the importance of a functional pyrR regulatory RNA element in growth, we performed growth experiments with mutants M1, M2 and M3 in the absence and presence of uracil. These data suggest that a functional pyrR does not appear to be absolutely necessary. For instance, while M1 may have a slight growth defect when cultured in the absence of uracil, M2 has no growth defect in the presence or absence of uracil (Fig 8A-B). Although both mutations result in constitutive expression of the *pyr* operon, mutation M1 leads to higher expression (Fig 7B) indicating that overexpression of the pyr genes may result in accumulation of end products that are detrimental to the cell. Alternatively, the M2 pyrR RNA element can still bind excess UMP-bound PyrR (as its PyrR binding domain is intact) thus

reducing the effective concentration of UMP in the cell and thereby potential accumulation associated side-effects. Importantly, M3 has a severe growth defect compared to wild type in the absence of uracil (Fig 8A), which can be partially rescued upon addition of uracil (Fig 8B). This suggests that while a constitutive off-state is detrimental for the bacterium in the absence of uracil a constitutive on-state can be overcome, indicating that efficient transcriptional control may not be essential.

To determine whether we can manipulate the manner in which the pyrR RNA element affects growth, we determined growth in the presence of 5-Fluoroorotic acid (5-FOA), a pyrimidine analog. 5-FOA is converted into 5-Fluorouracil (5-FU) a potent inhibitor of thymidylate synthetase, whose activity is essential for DNA replication and repair [45]. Additionally, 5-FU competes with UMP for interacting with the PyrR protein [46]. 5-FU can thus work as a decoy, signaling that UMP is present in the cell; triggering the formation of a terminator and reducing expression of the *pyr* operon. The wild type strain displayed a severe growth defect in the presence of 5-FOA (Fig 8C&D), while M1 (which should not interact with PyrR and should thus be largely insensitive to the presence of 5-FOA) displayed a much smaller growth defect (Fig 8C&D). In addition, M2, which constitutively over expresses the *pyr* operon, is also less sensitive to 5-FOA then wild type (Fig 8C&D). Thus, the mutations we introduced into the pyrR RNA element affect the secondary structure in the manner that we intended, and can have far reaching regulatory and fitness effects. Importantly, it shows that a drug targeted against the secondary structure can directly manipulate and severely hamper growth.

A remaining key question is the importance of RNA regulatory elements in colonization and the induction of disease. Somewhat surprisingly our *in vitro* growth curves suggest that constitutive expression of the *pyr* operon (M1) and constitutive overexpression (M2) is not all that detrimental to growth, indicating that efficient regulation is not critical. To assess the effect of loss of regulation on bacterial fitness *in vivo*, the pyrR RNA element mutants were tested in 1x1 competition assays (mutant vs. wild type) in a mouse infection model (Fig 9). While fitness for all three mutants is similar to the unmodified strain *in vitro* in the presence of uracil, M1 and M3 are unable to colonize and survive in the mouse nasopharynx, or infect and survive in the lung and transition and survive in the blood (Fig 9A&C). M2 has less of a defect *in vivo*, but still has a significantly diminished ability to infect and survive in the lung (Fig 9B). These results indicate that efficient regulation of the *pyr* operon *in vivo* is critical for growth and survival of *S. pneumoniae* within the host. While we had previously shown that genes in the *pyr* operon are important *in vivo* [44], the regulatory findings in this project take our understanding a step further and, importantly, in combination with the findings that 5-FOA can efficiently interact with the RNA regulatory element, suggests that it is feasible to modulate *in vivo* fitness and thereby virulence by targeting such regulatory elements.

Towards comprehensive transcriptional landscape reconstructions and highly targeted regulatory RNA element inhibitors.

With the advent of deep sequencing technologies, our understanding of prokaryotic transcriptional dynamics is rapidly advancing [47] and underlining that bacterial transcriptomes are not as simple as previously thought. Analysis of the *S. pneumoniae* TIGR4 transcriptome using three different sequencing techniques (RNA-Seq, term-seq, and 5'end-Seq) has led to a

306 comprehensive mapping of its transcriptional landscape. Besides identifying 1597 TSSs and
307 1330 sense and 236 antisense TTSs, we uncovered a complex operon structure, which has also
308 been found in *E. coli* [4]. Importantly, such complexity likely allows for environment-dependent
309 modulation of gene expression producing variable transcripts in response to varying conditions,
310 which we illustrated here through analyses of a 9-gene complex operon and the mal regulon (Fig
311 4). Additionally, similar environment-dependent versatile operon behavior has been observed in
312 *E. coli* [4] and to a lesser extent in *Mycoplasma pneumoniae* [48]. This means that our
313 understanding is shifting dramatically and it is thus becoming clear that operons in bacteria
314 should be seen as adaptable structures that can significantly increase the regulatory capacity of
315 the transcriptome by responding to environmental changes in a highly specific manner.
316
317 Another central aspect of our approach is the identification of putative 5'-UTR structured
318 regulatory elements. Riboswitches and other untranslated regulatory elements (binding sites for
319 small regulatory RNAs) are important bacterial RNA elements that are thought to regulate up to
320 2% of bacterial genes [7,8]. However, the discovery of new regulators is difficult when relying
321 solely on computational methodology and sequence conservation [49]. Here we show that
322 through term-seq [32] it is possible to identify such RNA elements on a genome-wide scale and
323 by combining it with RNA-Seq performed in different conditions transcriptional phenotypes can
324 be directly linked to the RNA element. This strategy thus makes it possible to screen for
325 regulatory RNA elements in retrospect by making use of already existing or newly generated
326 RNA-Seq data.
327

328 Importantly, besides the ability to re-construct an organism's intricate transcriptional landscape
329 we show that there is also a direct application of our multi-sequencing approach, namely the
330 ability to inhibit operons and/or pathways with specific chemicals or drugs that target the RNA
331 regulatory element. We show that this is possible for the pyrR RNA element, a regulatory
332 element that is important for pneumococcal growth and virulence, which means that this
333 regulatory element could be a potential antimicrobial drug target. This idea is further
334 strengthened by the fact that *S. pneumoniae* displays a growth defect in the presence of 5-FOA,
335 which directly relates to misregulation of pyrR RNA confirming its drug-able potential.
336
337 We believe that the presented multi-omics sequencing strategy brings a global understanding of
338 regulation in *S. pneumoniae* significantly closer, and because the approach is easily transferable
339 to other species, it will enable species-wide comparisons for conservation of operon structure and
340 regulatory elements. In addition, such detailed regulatory understanding creates new regulatory
341 control tools for synthetic biology purposes. Moreover, the combination with for instance *in vivo*
342 experiments shows that it is a realistic goal to design or select specific compounds that target
343 ribo-regulators in order to mitigate virulence or antibiotic resistance.
344
345 **Methods:**
346 *Culture conditions and sample collection*
347 For RNAtag-Seq, term-seq and 5'end-Seq library preparation, *Streptococcus pneumoniae* TIGR4
348 (T4) was cultured in rich media (SDMM) to mid-log phase (OD_{600} = 0.4). Cultures were diluted
349 to an OD_{600} of 0.05 in fresh media, grown for one doubling (T_0). At 0 min (T_0) and after 30 min
350 of growth (T_{30}) 10ml culture was harvested by means of centrifugation (4000 rpm, 7 min at 4^{o}C)

351 followed by flash freezing in a dry-ice ethanol bath and storage at -80^{o}C until RNA extraction.
352 Sample collection was performed in four biological replicates and total RNA was isolated using
353 an RNeasy Mini kit (Qiagen). For qRT-PCR analyses, T4 was cultured in SDMM to mid-log
354 phase (OD_{600} = 0.4) and after centrifugation cultures were washed with 1X PBS and diluted to an
355 OD_{600} of 0.003 in appropriate media. Cultures were harvested at mid-log followed by RNA
356 extraction as described above.

357
358 *5'end-Seq library preparation*

359 5'end-Seq libraries were generated by dividing the total RNA into 5' polyphosphate treated
360 (Processed) and untreated (Non-Processed) samples that were subsequently processed and
361 sequenced according to protocols described in Wurtzel et al., 2012 and [50] with few
362 modifications. See supplemental methods for a detailed protocol.

363
364 *RNA-Seq library preparation*

365 RNA-Seq libraries were generated by using the RNAtag-Seq protocol [31,51]. Briefly, 400 ng
366 RNA was fragmented in FastAP buffer, DNase-treated with Turbo DNase, 5'-dephosphorylated
367 using FastAP. Barcoded RNA adapters were then ligated to the 3' terminus, samples from
368 different conditions were pooled and ribosomal RNA was depleted using the Ribo-zero rRNA
369 removal kit. Illumina cDNA sequencing libraries were generated by first-strand cDNA synthesis,
370 3' linker ligation and PCR with 17 cycles. The final concentration and size distribution were
371 determined with the Qubit dsDNA BR Assay kit and the dsDNA D1000 Tapestation kit,
372 respectively.

373
374 *term-seq library preparation*

term-seq libraries were generated as previously described [32] with few modifications. 2 µg total RNA was depleted of genomic DNA using Turbo DNase, 5' dephosphorylated, ligated to barcoded RNA adapters at the 3' terminus and fragmented in fragmentation buffer. Barcoded and fragmented RNA from different conditions were pooled and ribosomal RNA was depleted using Ribo-zero. cDNA libraries were generated by first strand cDNA synthesis and RNA template was degraded as mentioned in the 5'end-Seq library preparation. Second 3' linker was ligated and PCR amplified for 17 cycles. All four library preparations (RNAtag-Seq, term-seq, 5'end-Seq processed and 5'end-Seq unprocessed) were pooled according to the method of preparation and sequenced at high depth (8.5 million reads/sample) on an Illumina NextSeq500.

Read processing and mapping

The sequencing reads from the 5' end-Seq sequencing, 3' end sequencing (term-seq), and RNAtag-Seq were processed and mapped to the *S. pneumoniae* TIGR4 (NC_003028.3) genome using the in-house developed Aerobio pipeline. Aerobio runs the processing and mapping in two phases. Phase 0 employs bcl2fastq to convert BCL to fastq files, quality control and de-multiplexing and compilation of the reads based on the sample conditions. Phase 1 maps the de-multiplexed reads against the genome, under default parameters, using Bowtie2 [52] and streams the output to SAMtools [53] to generate sorted and indexed BAM files for each sample.

in silico prediction of transcription start sites (TSSs) and transcription termination sites (TTSs)

Perl code from [32] was adapted to estimate the number of reads mapped at each nucleotide from the 5' end, 3' end, and RNA sequencing runs. With the nucleotide level coverage data calculated from the 5' end-Seq, regions up to 500 nucleotides upstream of the translational start sites described in the annotated TIGR4 genome (NC_003028.3) were scanned for mapped reads with

399 a minimum coverage of 2 and a Processed/non-Processed ratio of 1 as in [32]. When multiple
400 putative TSSs were identified in a 5' UTR, the one with the highest Processed/non-Processed
401 ratio was assigned as the TSS for the downstream gene. Similar to the identification of the TSSs,
402 TTSs were identified by scanning up to 150 nucleotides downstream of the translational stop site
403 for mapped 3' end reads with a minimum coverage of 2 in at least 4 replicates, out of the 12 total
404 datasets. The position with the highest coverage was considered the most likely TTS for a gene.
405
406 *Identification of transcript boundaries and operon structures in the genome*
407 BAM files of the mapped RNAtag-Seq reads were analyzed using Rockhopper [33] to predict
408 transcript boundaries and group genes into operons. Predicted operons were compared with the
409 genome based predictions listed in the Database of Prokaryotic Operons [5,34,54], and
410 complexity in the operon structure was characterized by surveying the number of internal TSSs
411 and TTSs similar to [4].
412
413 *Identification of candidate regulatory elements*
414 Once the TSSs were identified, 5'-UTR regions with a length of at least 70 nucleotides were
415 scanned for mapped 3' end sequencing reads with a minimum coverage of 2 to identify putative
416 early terminators. 5'-UTR regions with a predicted early TTS were binned as candidate
417 regulatory elements. The nucleotide sequence for each candidate element was obtained and
418 folded using RNAFold [55]. Secondary structures and free energy values were compiled for each
419 candidate. Putative candidates were compared to known bacterial non-coding RNAs described in
420 Rfam [56,57] that were identified in the genome by the cmsearch function of Infernal 1.1 [37].
421

The response of candidate regulatory elements to different media conditions were assessed by calculating the RNA-Seq coverage in both the regulatory element and the regulated gene. Read-through was calculated for each of the candidates as described previously [32]. Briefly, read-through is the ratio (denoted in percentage) of the average coverage across the gene to that of the 5'-UTR identified here. The greater the read-through, the higher the expression of the gene with respect to the 5'-UTR. That is, if the regulator reduced the expression of the gene, read-through would be small. If the regulator turned on gene expression in response to certain conditions, the read-through would be large.

Conservation of the candidate regulatory elements in Streptococcus pneumoniae

A local BLAST [58] database was generated with the genomes of 30 *S. pneumoniae* strains available in Refseq 77 [59] and 350 strains from [35]. Each of the candidate regulators identified in the genome of TIGR4 was BLASTed against this database, and hits in the other genomes were extracted and aligned using MAFFT version 7 [60]. The degree of conservation across the 380 genomes was determined by surveying each candidate cluster post filtering to remove sequences that were less than 70% in length of the query and with e-values greater than $1x10^{-4}$. The candidates were also screened for overlap with previously published small RNAs identified in *S. pneumoniae* [13,15].

Expression analysis using qPCR

RNA was isolated from cultures using the Qiagen RNeasy kit (Qiagen). DNase treated RNA was used to generate cDNA with iScript reverse transcriptase supermix for RT-qPCR (BioRad). Quantitative PCR was performed using a Bio-Rad MyiQ. Each sample was normalized against

445 the 50S ribosomal gene, SP2204 and were measured in biological replicate and technical
446 triplicates. No-reverse transcriptase and no-template controls were included for all samples.
447
448 *pyrR RNA mutant growth assays*
449 Wild type and pyrR RNA mutants of T4 were grown for 2 hours and diluted to an OD_{600} of 0.015
450 in fresh media, with varying concentrations of uracil and/or 5-FOA. Growth assays were
451 performed in 96-well plates for 16 hours by taking OD600 measurements every half hour using a
452 Tecan Infiniti Pro plate reader (Tecan). Growth assays were performed no less than two times.
453
454 *In vivo pyrR mutant fitness determination*
455 1 x 1 competition experiments were performed with pyrR RNA mutants (M1 to M3) that were
456 competed against the wild-type strain after which bacterial fitness was calculated as previously
457 described [44] with a few modifications. Lung removal and homogenization (in 10 mL 1 x PBS),
458 blood collection (100 uL) and nasopharnyx lavage (with 1 ml 1X PBS) were perfromed on all
459 animals 24 hours post infection, with the exception of pyrR M3, which due to the large fitness
460 defect were harvested at 6 hours post infection.
461
462 **Acknowledgements:** We would like to thank Jon Anthony for sequence data processing on the
463 Aerobio platform, Charles S. Hoffman for generous gift of 5-FOA that we used in this study and
464 Daniel Dar for discussion and helpful suggestions. The sequencing datasets generated during the
465 current study are available in the Sequence Read Archive (SRP136114). This work is supported
466 by NIH grant R01GM115931 to MMM and TvO, and R01AI110724 and U01AI124302 to TvO.
467

Author contributions. MMM and TvO devised the study. MMM, TvO, IW and ZZ designed the experiments, IW and ZZ generated the sequencing data, IW and AH performed *in vitro* experiments and AH performed *in vivo* experiments. IW, ZZ and NR performed RNA-Seq data analysis, NR performed term-seq and 5' end-Seq data analyses and IW, NR, MMM and TvO wrote the manuscript.

References

1. Jacob F, Manod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961;3: 318–356. doi:10.1016/S0022-2836(61)80072-7

2. Homuth G, Mogk A, Schumann W. Post-transcriptional regulation of the *Bacillus subtilis* dnaK operon. Molecular Microbiology. 1999;32: 1183–1197.

3. de Saizieu A, Certa U, Warrington J, Gray C, Keck W, Mous J. Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. Nat Biotech. 1998;16: 45–48. doi:10.1038/nbt0198-45

4. Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, et al. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. mBio. 2014;5: e01442–14. doi:10.1128/mBio.01442-14

5. Mao X, Ma Q, Liu B, Chen X, Zhang H, Xu Y. Revisiting operons: an analysis of the landscape of transcriptional units in *E. coli*. BMC Bioinformatics. 6 ed. 2015;16: 356. doi:10.1186/s12859-015-0805-8

6. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature. 2010;464: 250–255. doi:10.1038/nature08756

7. Mandal M, Boese B, Barrick J, Winkler W, Breaker R. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. Cell. 2003;113: 577–586.

8. Waters LS, Storz G. Regulatory RNAs in Bacteria. Cell. 2009;136: 615–628. doi:10.1016/j.cell.2009.01.043

9. Saberi F, Kamali M, Najafi A, Yazdanparast A, Moghaddam MM. Natural antisense RNAs

as mRNA regulatory elements in bacteria: a review on function and applications. Cell Mol Biol Lett. BioMed Central; 2016;21: 6. doi:10.1186/s11658-016-0007-z

10. Ignatov D, Johansson J. RNA-mediated signal perception in pathogenic bacteria. WIREs RNA. 2017;8. doi:10.1002/wrna.1429

11. Eyraud A, Tattevin P, Chabelskaya S, Felden B. A small RNA controls a protein regulator involved in antibiotic resistance in *Staphylococcus aureus.* Nucleic Acids Res. 2014;42: 4892–4905. doi:10.1093/nar/gku149

12. Vanderpool CK, Balasubramanian D, Lloyd CR. Dual-function RNA regulators in bacteria. Biochimie. 2011;93: 1943–1949. doi:10.1016/j.biochi.2011.07.016

13. Acebo P, Martin-Galiano AJ, Navarro S, Zaballos A, Amblar M. Identification of 88 regulatory small RNAs in the TIGR4 strain of the human pathogen *Streptococcus pneumoniae*. RNA. 2012;18: 530–546. doi:10.1261/rna.027359.111

14. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011;471: 602–607. doi:10.1038/nature09886

15. Mann B, van Opijnen T, Wang J, Obert C, Wang Y-D, Carter R, et al. Control of Virulence by Small RNAs in *Streptococcus pneumoniae*. Cossart P, editor. PLoS Pathog. 2012;8: e1002788. doi:10.1371/journal.ppat.1002788.s001

16. Pichon C, Merle du L, Caliot M-E, Trieu-Cuot P, Le Bouguenec C. An in silico model for identification of small RNAs in whole bacterial genomes: characterization of antisense RNAs in pathogenic *Escherichia coli* and *Streptococcus agalactiae* strains. Nucleic Acids Res. 2012;40: 2846–2861. doi:10.1093/nar/gkr1141

17. Rosinski-Chupin I, Sauvage E, Sismeiro O, Villain A, Da Cunha V, Caliot M-E, et al. Single nucleotide resolution RNA-seq uncovers new regulatory mechanisms in the opportunistic pathogen *Streptococcus agalactiae*. BMC Genomics. 2015;16: 2906–15. doi:10.1186/s12864-015-1583-4

18. Tesorero RA, Yu N, Wright JO, Svencionis JP, Cheng Q, Kim J-H, et al. Novel regulatory small RNAs in *Streptococcus pyogenes*. PLoS ONE. 2013;8: e64021. doi:10.1371/journal.pone.0064021

19. Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Bécavin C, et al. Comparative transcriptomics of pathogenic and non-pathogenic Listeria species. Molecular Systems Biology. 2012;8: 583. doi:10.1038/msb.2012.11

20. Papenfort K, Vogel J. Regulatory RNA in Bacterial Pathogens. Cell Host Microbe. Elsevier Inc; 2010;8: 116–127. doi:10.1016/j.chom.2010.06.008

21. Caldelari I, Chao Y, Romby P, Vogel J. RNA-Mediated Regulation in Pathogenic Bacteria. Cold Spring Harbor Perspectives in Medicine. 2013;3: a010298–a010298. doi:10.1101/cshperspect.a010298

22. Blount KF, Wang JX, Lim J, Sudarsan N, Breaker RR. Antibacterial lysine analogs that target lysine riboswitches. Nat Chem Biol. 2006;3: 44–49. doi:10.1038/nchembio842

23. Lunse CE, Schuller A, Mayer G. The promise of riboswitches as potential antibacterial drug targets. Int J Med Microbiol. 2014;304: 79–92. doi:10.1016/j.ijmm.2013.09.002

24. Mulhbacher J, St-Pierre P, Lafontaine DA. Therapeutic applications of ribozymes and riboswitches. Curr Opin Pharmacol. Elsevier Ltd; 2010;10: 551–556. doi:10.1016/j.coph.2010.07.002

25. Howe JA, Wang H, Fischmann TO, Balibar CJ, Xiao L, Galgoci AM, et al. Selective small-molecule inhibition of an RNA structural element. Nature. 2015;526: 672–677. doi:10.1038/nature15542

26. CDC. Antibiotic Resistance Threats in the United States, 2013. CDC report. 2013;: 1–114.

27. Henriques-Normark B, Tuomanen EI. The pneumococcus: epidemiology, microbiology, and pathogenesis. Cold Spring Harbor Perspectives in Medicine. 2013;3. doi:10.1101/cshperspect.a010215

28. WHO. The World Heath Report 2007. World Health Organization; 2007. pp. 1–96.

29. Wu Z, Wu C, Shao J, Zhu Z, Wang W, Zhang W, et al. The *Streptococcus suis* transcriptional landscape reveals adaptation mechanisms in pig blood and cerebrospinal fluid. RNA. 2014;20: 882–898. doi:10.1261/rna.041822.113

30. Kumar R, Shah P, Swiatlo E, Burgess SC, Lawrence ML, Nanduri B. Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays. BMC Genomics. 2010;11: 350. doi:10.1186/1471-2164-11-350

31. Shishkin AA, Giannoukos G, Kucukural A, Ciulla D, Busby M, Surka C, et al. Simultaneous generation of many RNA-seq libraries in a single reaction. Nat Meth. 2015;12: 323–325. doi:10.1038/nmeth.3313

32. Dar D, Shamir M, Mellin JR, Koutero M, Stern-Ginossar N, Cossart P, et al. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science. 2016;352: 187–187. doi:10.1126/science.aad9822

33. Tjaden B. *De novo* assembly of bacterial transcriptomes from RNA-seq data. Genome Biol. 2015;16: 1. doi:10.1186/s13059-014-0572-2

34. Mao F, Dam P, Chou J, Olman V, Xu Y. DOOR: a database for prokaryotic operons. Nucleic Acids Res. 2009;37: D459.

35. Cremers AJH, Mobegi FM, de Jonge MI, van Hijum SAFT, Meis JF, Hermans PWM, et al. The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. Sci Rep. Nature Publishing Group; 2015;5: 14952. doi:10.1038/srep14952

36. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution. 2016;33: 1870–1874. doi:10.1093/molbev/msw054

37. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29: 2933–2935. doi:10.1093/bioinformatics/btt509

38. Winkler W, Nahvi A, Breaker R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature. 2002;419: 952–956.

39. Winkler W, Cohen-Chalamish S, Breaker R. An mRNA structure that controls gene expression by binding FMN. Proc Natl Acad Sci USA. 2002;99: 15908.

40. Switzer RL, Turner RJ, Lu Y. Regulation of the *Bacillus subtilis* pyrimidine biosynthetic operon by transcriptional attenuation: control of gene expression by an mRNA-binding protein. Prog Nucleic Acid Res Mol Biol. 1999;62: 329–367.

41. Turner RJ, Lu Y, Switzer RL. Regulation of the *Bacillus subtilis* pyrimidine biosynthetic (*pyr*) gene cluster by an autogenous transcriptional attenuation mechanism. J Bacteriol. 1994;176: 3708–3722.

42. Tomchick DR, Turner RJ, Switzer RL, Smith JL. Adaptation of an enzyme to regulatory function: structure of *Bacillus subtilis* PyrR, a *pyr* RNA-binding attenuation protein and uracil phosphoribosyltransferase. Structure. 1998;6: 337–350.

43. Lu Y, Turner RJ, Switzer RL. Function of RNA secondary structures in transcriptional attenuation of the *Bacillus subtilis pyr* operon. Proc Natl Acad Sci USA. 1996;93: 14462–14467.

44. van Opijnen T, Camilli A. A fine scale phenotype-genotype virulence map of a bacterial pathogen. Genome Research. 2012;22: 2541–2551. doi:10.1101/gr.137430.112

45. Longley DB, Harkin DP, Johnston PG. 5-fluorouracil: mechanisms of action and clinical strategies. Nature Reviews Cancer. 2003;3: 330–338. doi:10.1038/nrc1074

46. Ghode P, Ramachandran S, Bifani P, Sivaraman J. Structure and mapping of spontaneous mutational sites of PyrR from *Mycobacterium tuberculosis*. Biochemical and biophysical research communications. 2016;471: 409–415. doi:10.1016/j.bbrc.2016.02.071

600 47. James K, Cockell SJ, Zenkin N. Deep sequencing approaches for the analysis of
601 prokaryotic transcriptional boundaries and dynamics. Methods. 2017;120: 76–84.
602 doi:10.1016/j.ymeth.2017.04.016

603 48. Güell M, Van Noort V, Yus E, Chen W, Leigh-Bell J, Michalodimitrakis K, et al.
604 Transcriptome complexity in a genome-reduced bacterium. Science. 2009;326: 1268.

605 49. Breaker RR. Prospects for Riboswitch Discovery and Analysis. Molecular Cell. Elsevier
606 Inc; 2011;43: 867–879. doi:10.1016/j.molcel.2011.08.024

607 50. Shell SS, Chase MR, Ioerger TR, Fortune SM. RNA sequencing for transcript 5'-end
608 mapping in mycobacteria. Methods Mol Biol. 2015;1285: 31–45. doi:10.1007/978-1-4939-
609 2450-9_3

610 51. Jensen PA, Zhu Z, van Opijnen T. Antibiotics Disrupt Coordination between
611 Transcriptional and Phenotypic Stress Responses in Pathogenic Bacteria. CellReports.
612 2017;20: 1705–1716. doi:10.1016/j.celrep.2017.07.062

613 52. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment
614 of short DNA sequences to the human genome. Genome Biol. 2009;10: R25.
615 doi:10.1186/gb-2009-10-3-r25

616 53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
617 Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.
618 doi:10.1093/bioinformatics/btp352

619 54. Dam P, Olman V, Harris K, Su Z, Xu Y. Operon prediction using both genome-specific
620 and general genomic information. Nucleic Acids Res. 2006;35: 288–298.
621 doi:10.1093/nar/gkl1018

622 55. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al.
623 ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6: 26. doi:10.1186/1748-7188-6-26

624 56. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family
625 database. Nucleic Acids Res. 2003;31: 439–441. doi:10.1093/nar/gkg006

626 57. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0:
627 updates to the RNA families database. Nucleic Acids Res. 2015;43: D130–7.
628 doi:10.1093/nar/gku1063

629 58. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol
630 Biol. 1990;215: 403–410.

631 59. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference
632 sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and
633 functional annotation. Nucleic Acids Res. 2016;44: D733–45. doi:10.1093/nar/gkv1189

634 60. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
635 improvements in performance and usability. Molecular Biology and Evolution. 2013;30:
636 772–780. doi:10.1093/molbev/mst010

637

638
639

Figure legends

Figure 1. Schematic representation of the sequencing and data analysis methodology. (A) A description of the experimental pipeline. RNA-Seq (left column) and term-seq (middle column) libraries were prepared according to protocols described in [31,32]. 5'end-Seq libraries (right column) were generated by dividing the total RNA into 5' polyphosphate treated (Processed) and untreated (Non-Processed) samples and subsequently processed according to protocols described in [19,50]. White and grey lines correspond to RNA and cDNA, respectively and colored blocks represent unique sequence barcodes. Illumina sequencing adaptors with index barcodes are also indicated. **(B)** A brief description of the analysis pipeline. Raw reads from all three sequencing methodologies describes above were de-multiplexed and aligned to T4 (NC_003028.3). Based on the reads mapped, single nucleotide coverage was calculated. Coverage calculated for the 5' end-Seq was used to determine the transcription start sites. Coverage from term-seq was used to determine the transcription termination sites. RNA-Seq coverage was used to calculate the read-through across candidate 5' untranslated regions with early transcription terminators.

Figure 2. Genome-wide map of the *Streptococcus pneumoniae* TIGR4 transcriptional landscape. A map of all the transcriptional features identified. The internal band represents the 1597 transcription start sites (TSSs) in green; 565 putative candidate regulatory regions in black; 565 early transcription termination sites in the 5'-UTR in orange; 1330 sense transcription termination sites (TTS) in red; and 236 antisense transcription termination sites in blue. The outer band of the genome map represents the annotated operon structures that were classified according to their number of TSS and TTSs: 1) traditional operons consisting of multiple genes with a single TSS and a TTS (green); 2) multiTSS operons consisting of multiple genes with internal TSSs but one TTS (blue); 3) multiTTS operons consisting of multiple genes with a single TSS but multiple internal TTSs (red); and 4) complex operons consisting of multiple genes with multiple internal TSSs and TTSs (orange). To avoid clutter, simple operons consisting of a single gene with a single TTS and TSS are not shown.

Figure 3. Distribution of operon types in the genome and frequency of transcriptional features within non-traditional operons. (A) The pie chart describes the distribution of the

types of operons present in T4. A total of 474 multigene and 773 single gene operons were identified, which can be divided up in 62% simple operons (single gene transcriptional units with a single TSS and TTS; gray), 26% complex operons (multi-gene operons with multiple TSSs and TTSs; orange), 5% traditional operons (multi-gene operon with a single TSS and TTS; green), 4% multiTSS operons (blue), and 3% multiTTS operons (red). (**B)** The histogram describes the distribution of genes and transcriptional features in non-traditional operons, where gray represents the numbers of genes in the multigene operons, green represents the number of TSSs within operons, red represents the number of TTSs within operons, and overlapping numbers are shown in brown. Two-gene operons are found most frequently in the non-traditional operons with one internal TSS and TTS.

Figure 4. Variability in expression levels between genes in the same operon when grown in rich (SDMM) and poor (MCDM) media conditions. RNA-Seq coverage maps of complex operon/regulon including the TSSs in green and TTSs in red. Size of the transcriptional features represents log transformation of the Processed/Non-Processed ratio for TSSs and coverage for TTSs. (**A)** A complex 9-gene operon (SP1018–SP1026) encoding thymidine kinase, GNAT family N-acetyltransferase, peptide chain release factor 1, peptide chain release factor N(5)-glutamine methyltransferase, threonylcarbomyl-AMP synthase, N-acetyltransferase, serine hydroxymethyltransferase, nucleoid-associated protein, and Pneumococcal vaccine antigen A respectively. Genes SP1022-SP1026 are expressed to greater levels in MCDM than in SDMM unlike genes SP1018-SP1021. (**B)** The maltose regulon, an example of three operons of different complexities working together in response to maltose in the medium. Complex operon SP2108-SP2110 shows greater expression in MCDM than SDMM in comparison to the simple operon SP2111-SP2112 and multiTTS operon SP2106-SP2017

Figure 5. Validation of the regulatory activities of FMN and TPP riboswitches in different nutrient conditions. The relative expression and average RNA-Seq coverage of SP0178 **(A)** and SP0716 **(C)** increases in poor (MCDM) medium compared to rich (SDMM) medium, potentially compensating for the depletion of the specific ligand. Expression of SP0178 **(B)** and SP0716 **(D)** is reduced when the poor medium is supplemented with respective ligands thus confirming the regulatory activities of FMN and TPP riboswitches. (FMN- Riboflavin; TPP- Thiamine).

Figure 6. Mechanism and regulatory activity of pyrR regulatory RNA element in the presence and absence of uracil. (**A)** Schematic representation of the proposed mechanism of regulation of *pyr* operon by pyrR RNA element. In the presence of UMP, PyrR binds to the pyrR RNA and results in the formation of a premature terminator, disrupting the anti-terminator formed when UMP is low, resulting in transcription termination. **(B)** RNA-Seq coverage map across the *pyr* operon (SP1278-1276) showing premature transcription termination and consequently decreased expression of its genes downstream of the pyrR regulator when grown in defined medium (CDM) in the presence of uracil (yellow) compared to the absence of uracil (blue). TSSs are in green and the size represents the log transformation of the Processed/Non-Processed ratio and TTSs are in red and size represents the log transformation of the coverage. **(C)** qRT-PCR determining the expression of the first genes in the *pyr* operon (SP1278) in the presence and absence of uracil validates the RNA-Seq observation.

Figure 7. Structure and regulatory activity of pyrR RNA mutants. (**A)** Secondary structure of the *S. pneumoniae* pyrR RNA regulatory element in 'off' conformation. Boxed in red and yellow are bases that were deleted in M1 and M3 mutations, respectively. Bases boxed in green were replaced by indicated bases to make mutation M2. Highlighted in grey are nucleotides that would base pair to form the anti-terminator when the riboswitch is in the "on" conformation. (**B)** A representative qRT-PCR quantification of the expression of SP1278 transcript from pyrR RNA mutant strains cultured in defined medium with or without uracil, corresponding to the regulatory activity of pyrR RNA mutants. While the WTc (wild type with chloramphenicol resistance cassette) decreases expression in the presence of uracil, M1 is insensitive to the ligand. M2 and M3 result in either constitutive or reduced expression of the *pyr* operon. ~2 fold higher expression in M1 compared to WTc in the absence of uracil could be the result of endogenous uracil having a slight inhibiting effect on the wild type.

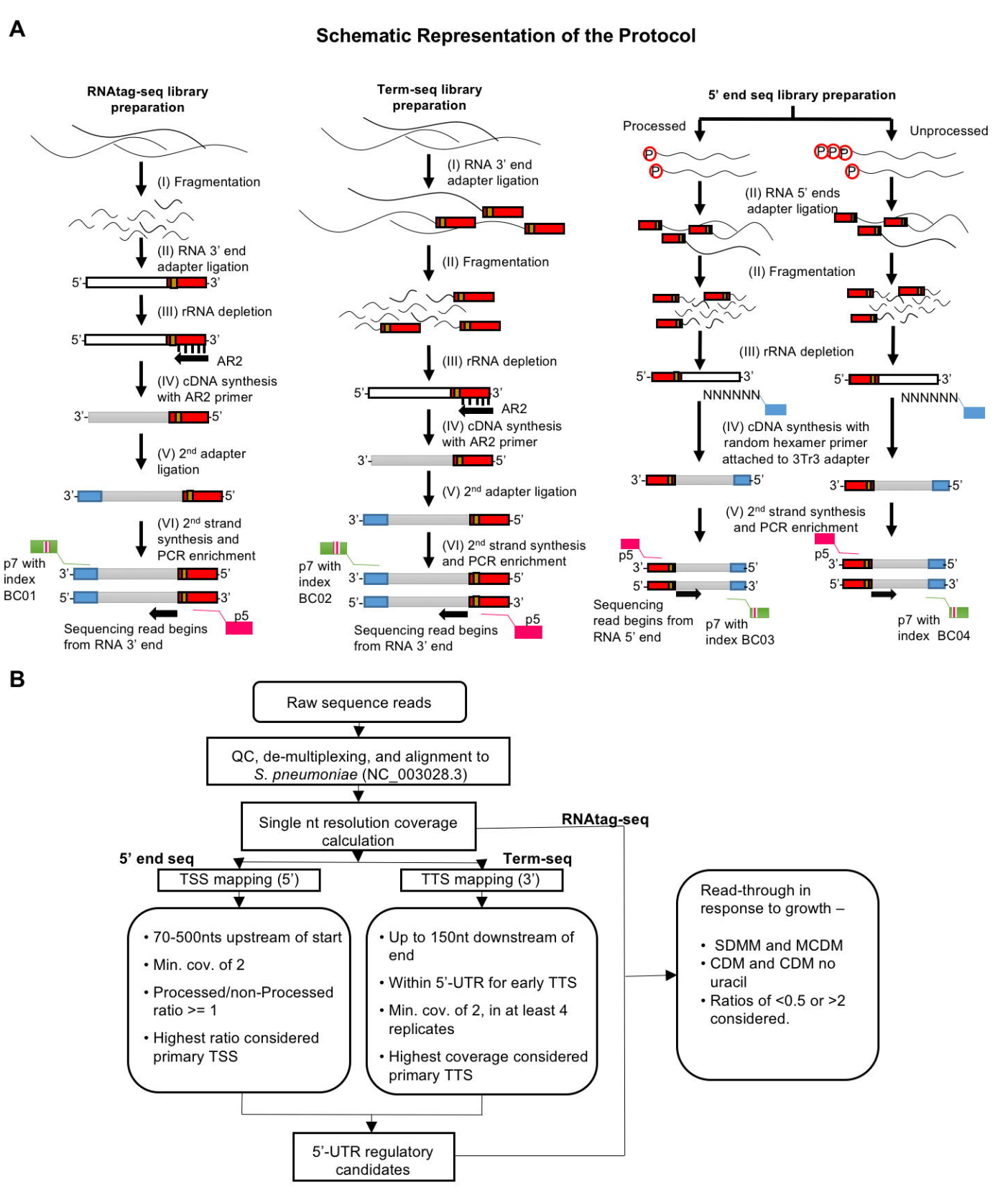
Figure 8. Regulation by pyrR regulatory element is important, but not essential for *in vitro* growth of *S. pneumoniae*. (A-B) Representative *in vitro* growth curves of mutants when cultured in defined media with (20 µg/ml) or without uracil. While mutant M2 (green) does not display a growth defect, mutants M1 (orange) and M3 (maroon) have growth defects that are

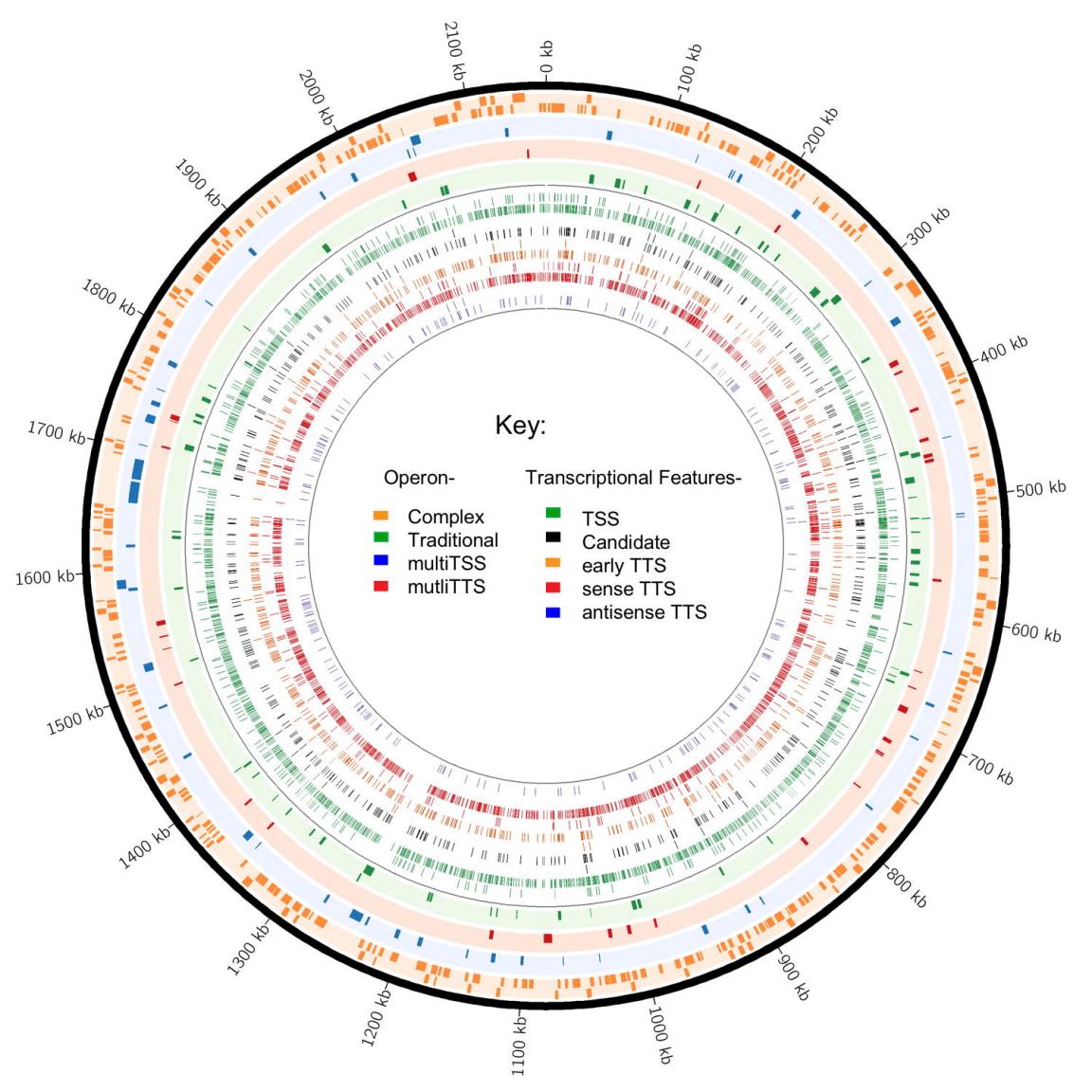
restored in the presence of uracil indicating that a functional pyrR RNA element is important, but not absolutely essential for *in vitro* growth of *S. pneumoniae*. WT (blue) doesn't have any growth defect in the tested conditions. **(C-D)** Representative *in vitro* growth curves of mutants cultured in media with (15 µg/ml) or without 5-FOA, a toxic uracil analog. All strains show varying degrees of defects in the tested conditions indicating that a drug targeted against the secondary structure can severely and specifically hamper growth. Mutant M3 was not included in this assay as it had a severe growth defect in the assay condition.

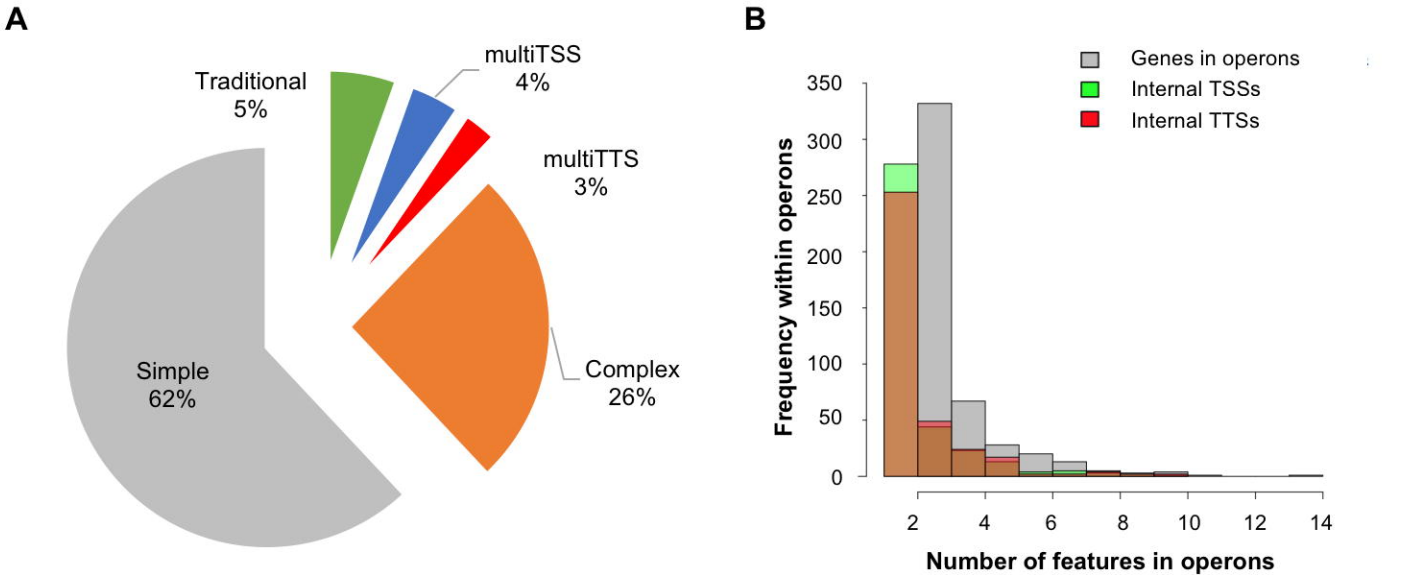
Figure 9. Regulation by the pyrR RNA element is crucial for *in vivo* survival and virulence of *S. pneumoniae*. 1x1 competition assays (mutant vs wild type T4) reveals fitness defects of pyrR RNA mutants in a mouse infection model. While M1 (A) and M3 (C) display severe defects in all the tested *in vivo* environments namely lung, blood and nasopharynx, M2 (B) has less of a defect. Significant change in fitness ($p<0.0125$) are indicated by asterisks (*). Each data point represents a single mouse.

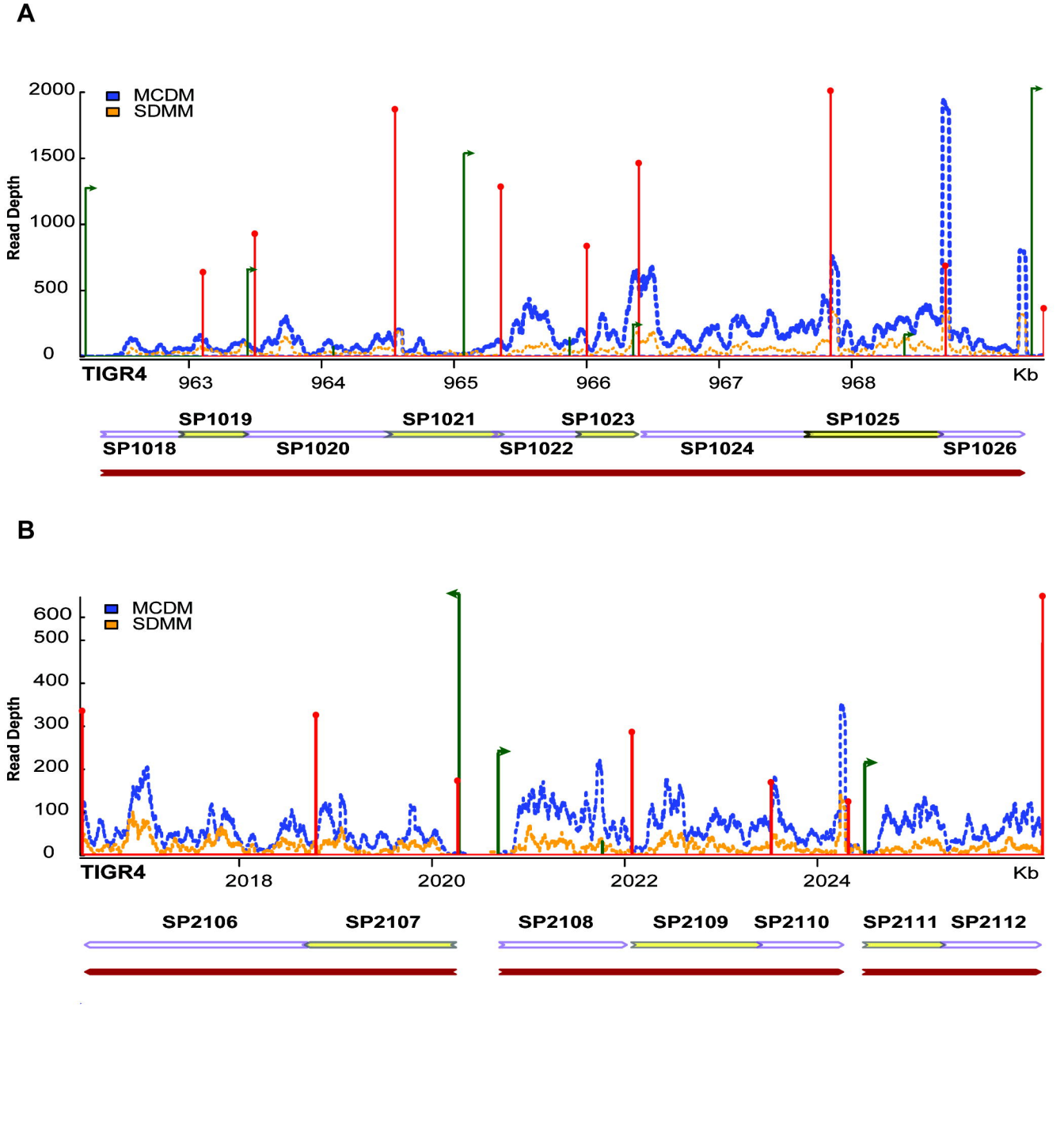
Supplemental figure 1. Distribution and conservation of the 565 putative regulatory candidates across 380 *S. pneumoniae* genomes. A. Frequency distribution of the candidates across the surveyed genomes. **B.** Conservation of the candidates as a measure of the mean p-Distance within each candidate cluster.

Supplemental figure 2. Validation of the regulatory activities of three putative 5'-UTR regulatory candidates in different nutrient conditions. The relative expression and average RNA-Seq coverage of SP1356 **(A)**, SP0240 **(B)** and SP1951 **(C)** increases in poor media (MCDM) compared to rich media (SDMM), potentially compensating for the depletion of the specific ligand.

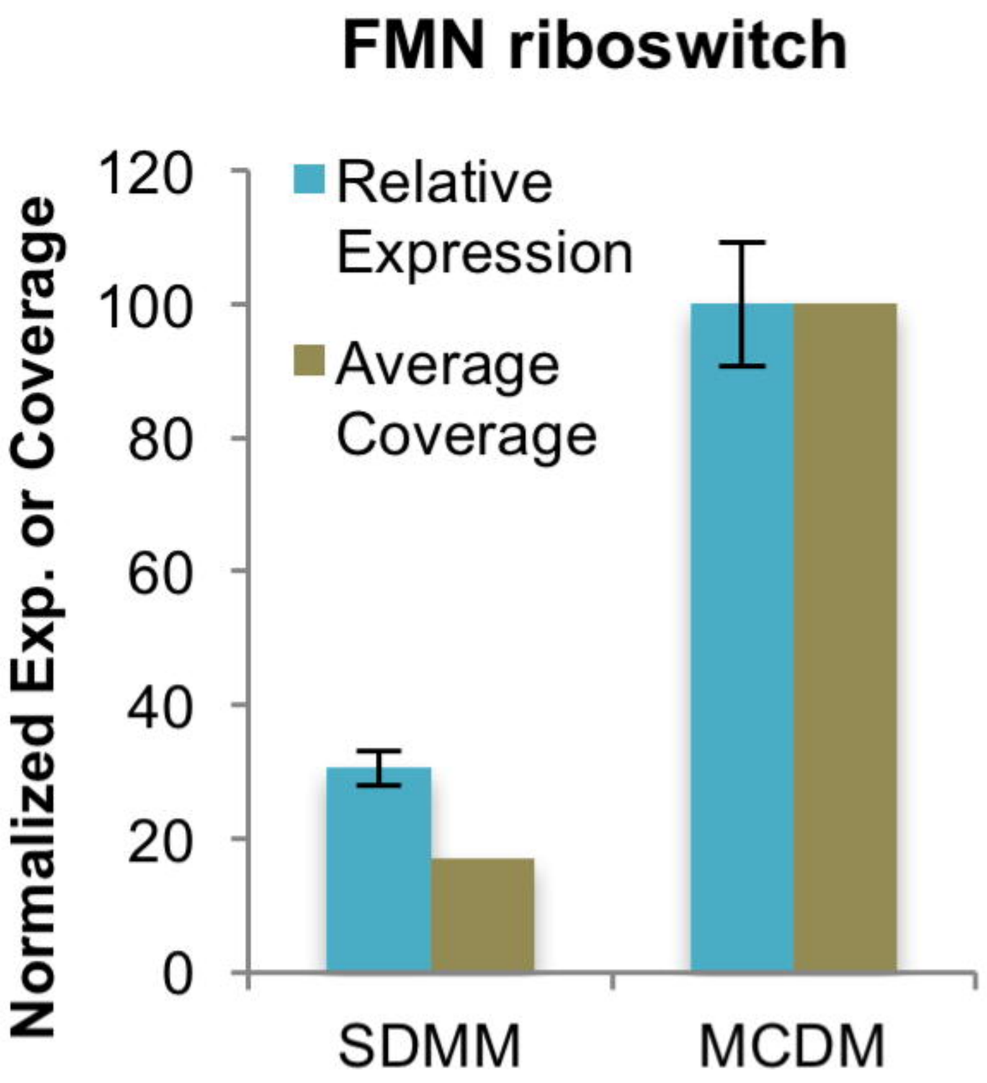

A
Schematic Representation of the Protocol
RNAtag-seq library preparation
(I) Fragmentation
(II) RNA 3' end adapter ligation
5'-3'
(III) rRNA depletion
AR2
(IV) cDNA synthesis with AR2 primer
3'-5'
(V) 2nd adapter ligation
(VI) 2nd strand synthesis and PCR enrichment
p7 with index BC01
Sequencing read begins from RNA 3' end
p5
Term-seq library preparation
(I) RNA 3' end adapter ligation
(II) Fragmentation
(III) rRNA depletion
AR2
(IV) cDNA synthesis with AR2 primer
(V) 2nd adapter ligation
(VI) 2nd strand synthesis and PCR enrichment
p7 with index BC02
Sequencing read begins from RNA 3' end
p5
5' end seq library preparation
Processed
Unprocessed
(II) RNA 5' ends adapter ligation
(II) Fragmentation
(III) rRNA depletion
NNNNNN
(IV) cDNA synthesis with random hexamer primer attached to 3Tr3 adapter
(V) 2nd strand synthesis and PCR enrichment
p5
Sequencing read begins from RNA 5' end
p7 with index BC03
p7 with index BC04
B
Raw sequence reads
QC, de-multiplexing, and alignment to S. pneumoniae (NC_003028.3)
Single nt resolution coverage calculation
RNAtag-seq
5' end seq
Term-seq
TSS mapping (5')
TTS mapping (3')
• 70-500nts upstream of start
• Min. cov. of 2
• Processed/non-Processed ratio >= 1
• Highest ratio considered primary TSS
• Up to 150nt downstream of end
• Within 5'-UTR for early TTS
• Min. cov. of 2, in at least 4 replicates
• Highest coverage considered primary TTS
Read-through in response to growth –
• SDMM and MCDM
• CDM and CDM no uracil
• Ratios of <0.5 or >2 considered.
5'-UTR regulatory candidates


Key:
Operon-
Complex
Traditional
multiTSS
mutliTTS
Transcriptional Features-
TSS
Candidate
early TTS
sense TTS
antisense TTS
0 kb
100 kb
200 kb
300 kb
400 kb
500 kb
600 kb
700 kb
800 kb
900 kb
1000 kb
1100 kb
1200 kb
1300 kb
1400 kb
1500 kb
1600 kb
1700 kb
1800 kb
1900 kb
2000 kb
2100 kb

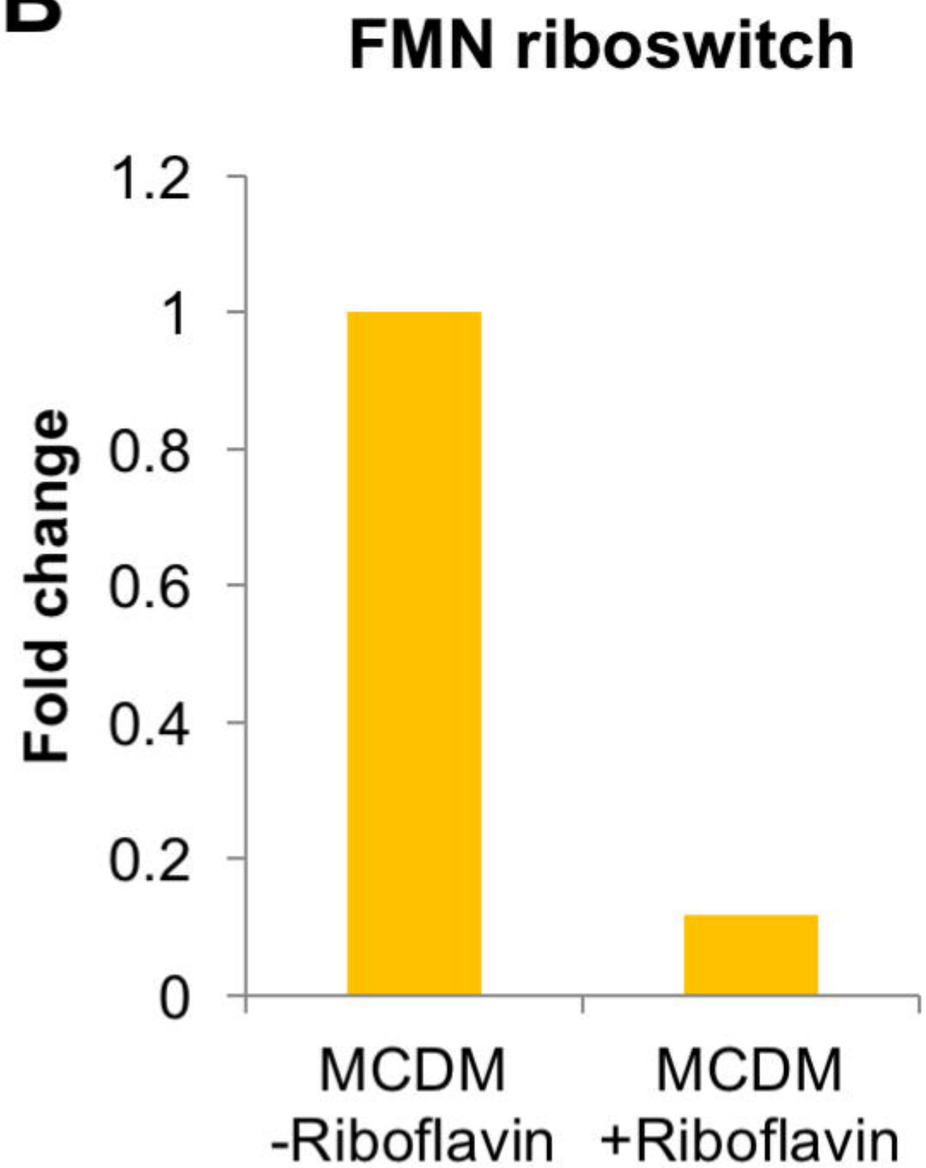

A
Traditional
5%
multiTSS
4%
multiTTS
3%
Complex
26%
Simple
62%
B
Genes in operons
Internal TSSs
Internal TTSs
Frequency within operons
0
50
100
150
200
250
300
350
2
4
6
8
10
12
14
Number of features in operons


A
MCDM
SDMM
Read Depth
2000
1500
1000
500
0
TIGR4
963
964
965
966
967
968
Kb
SP1019
SP1021
SP1023
SP1025
SP1018
SP1020
SP1022
SP1024
SP1026
B
MCDM
SDMM
Read Depth
600
500
400
300
200
100
0
TIGR4
2018
2020
2022
2024
Kb
SP2106
SP2107
SP2108
SP2109
SP2110
SP2111
SP2112

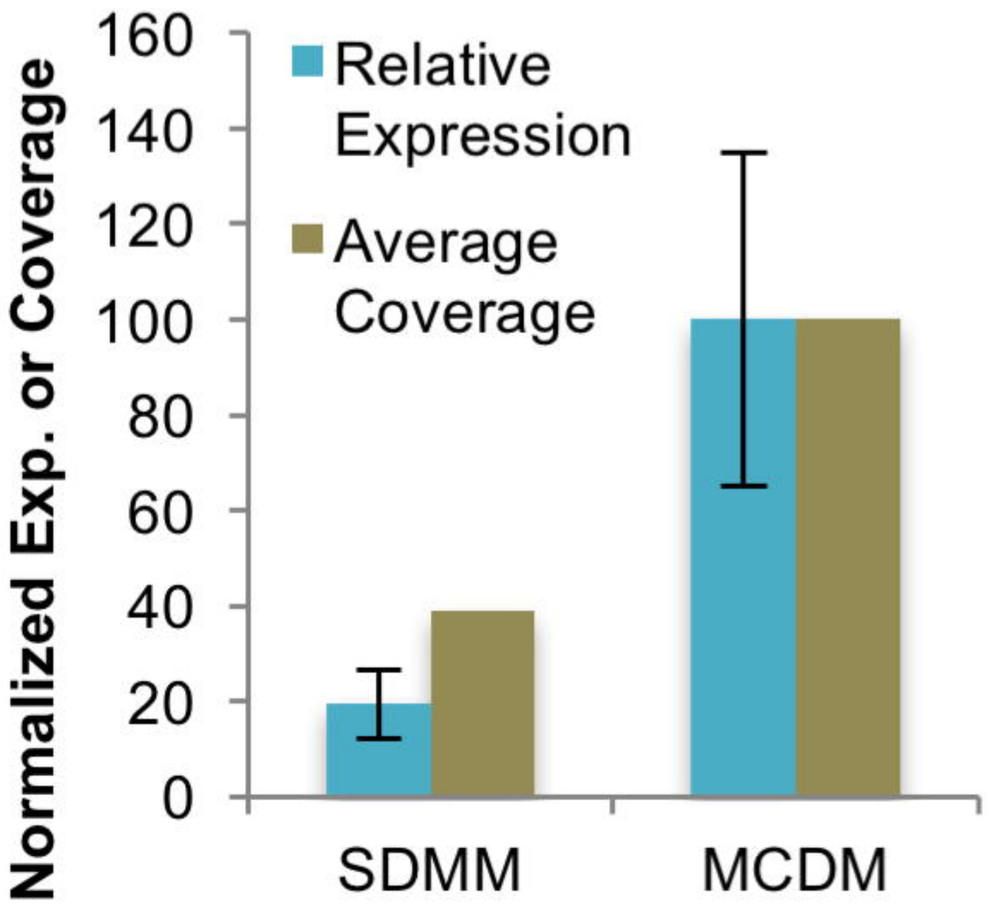
A



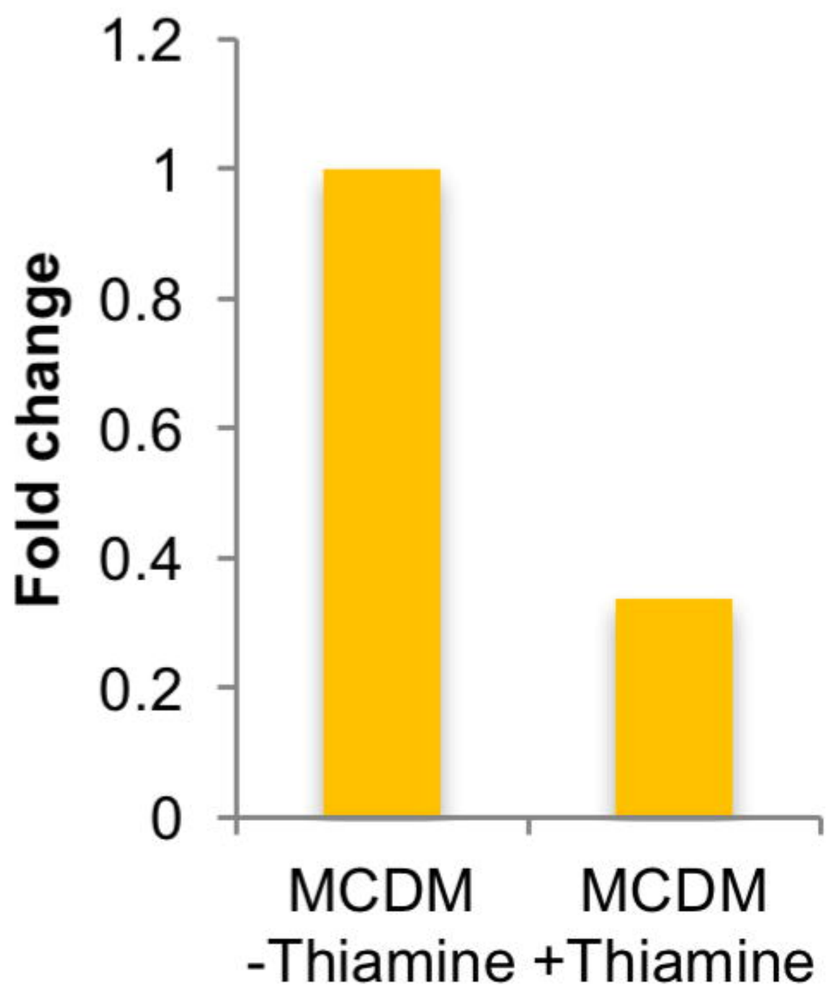
B

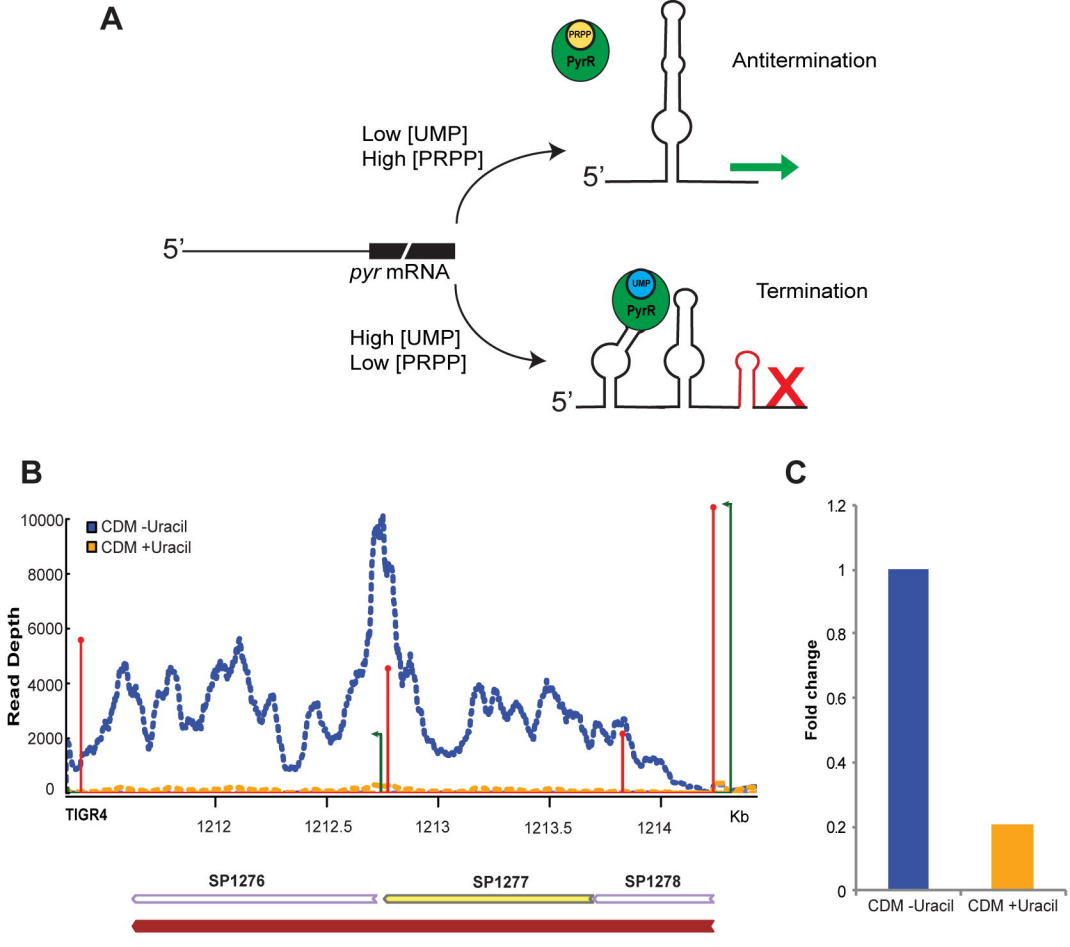


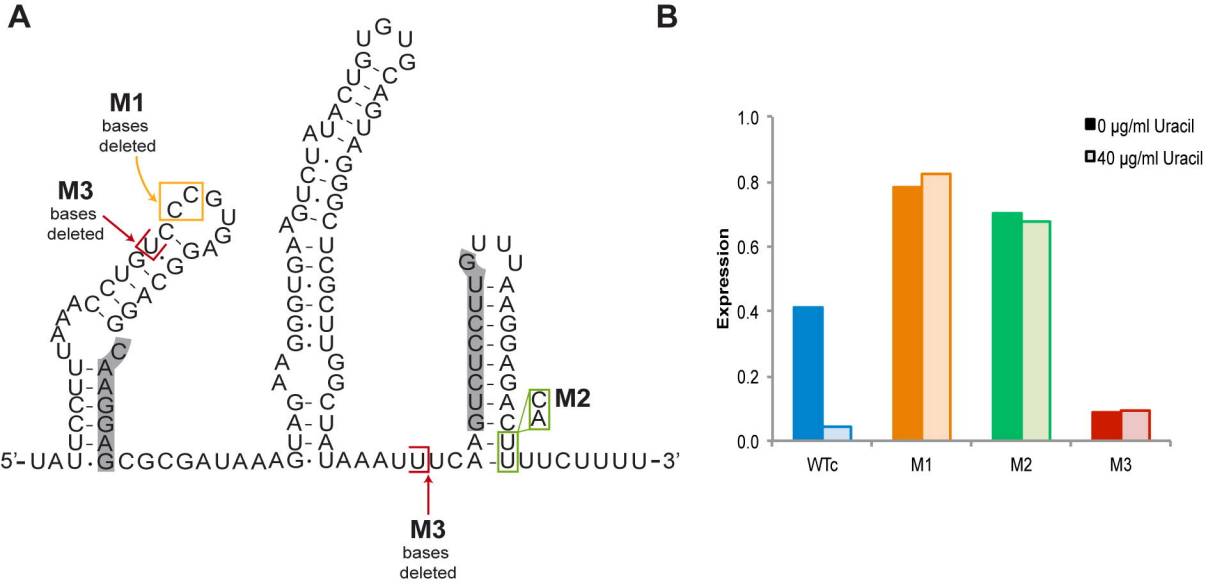
C

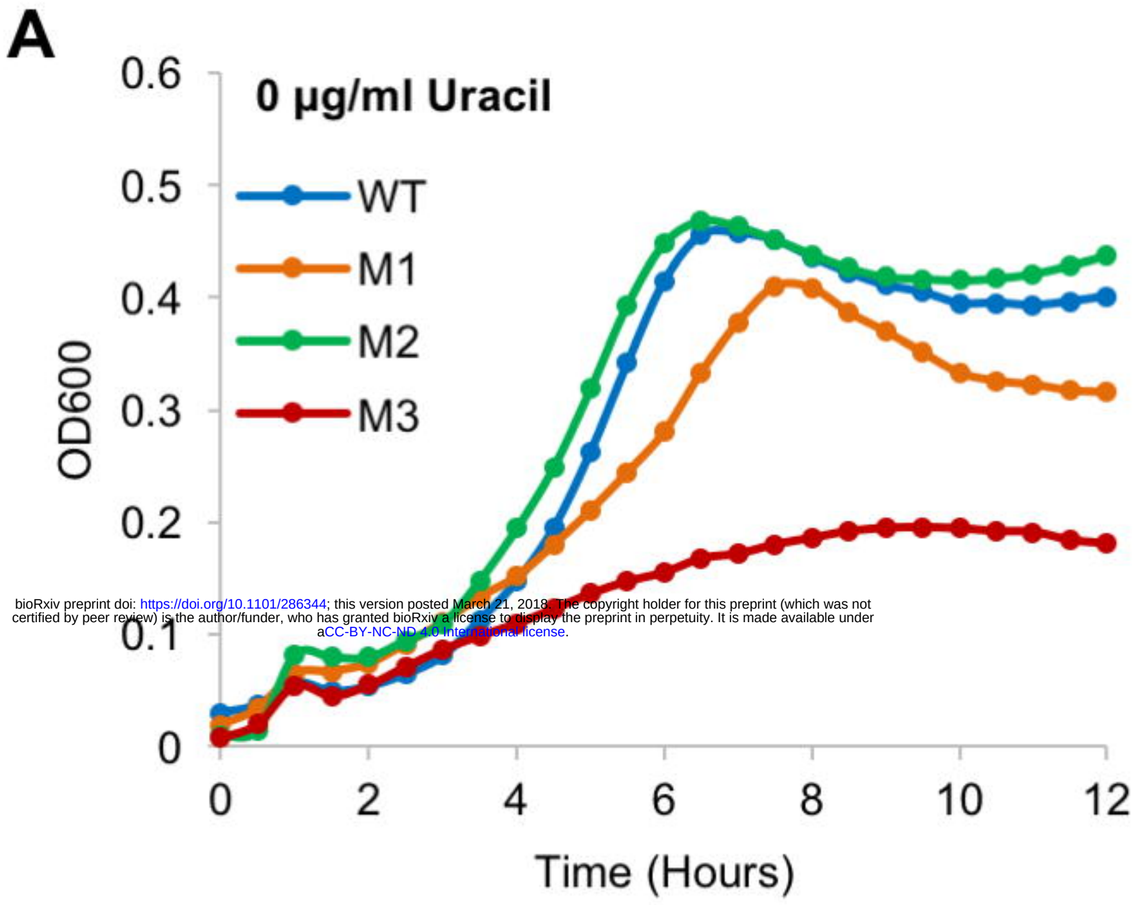


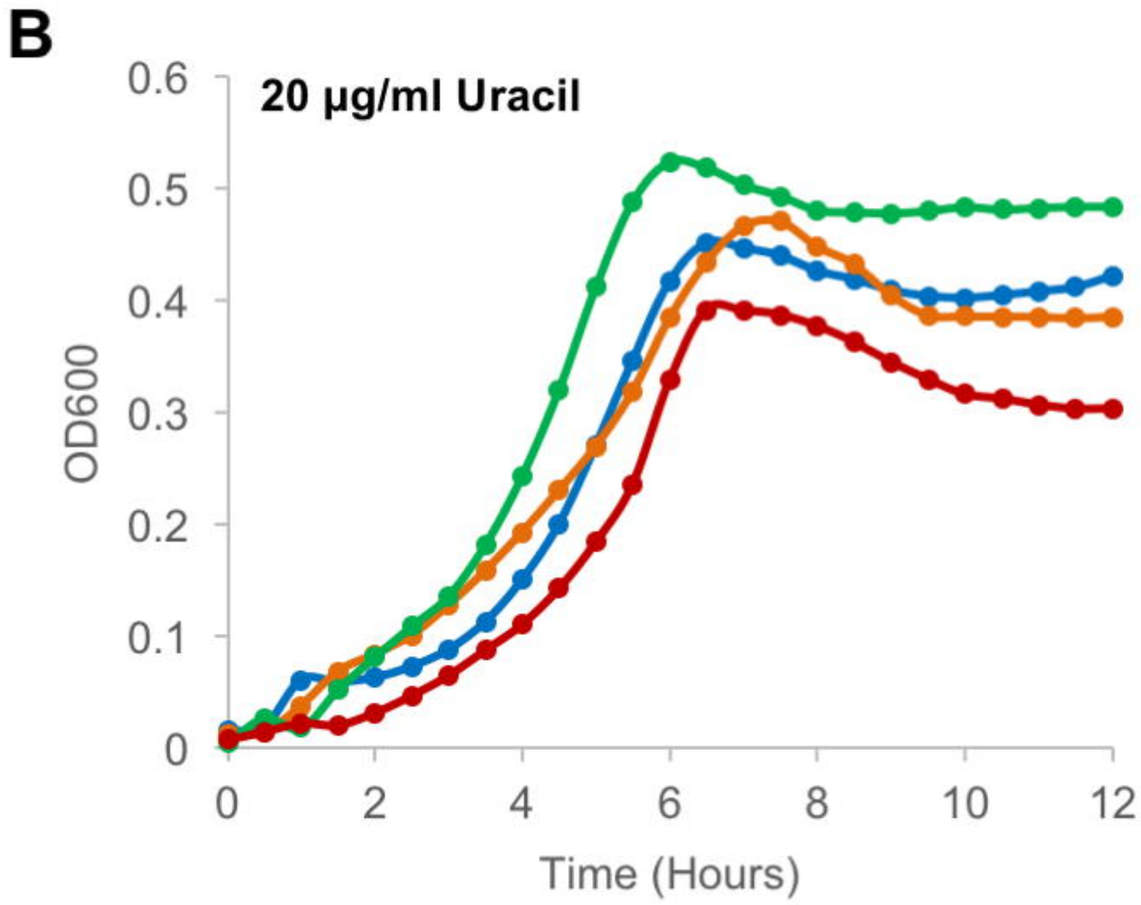
D

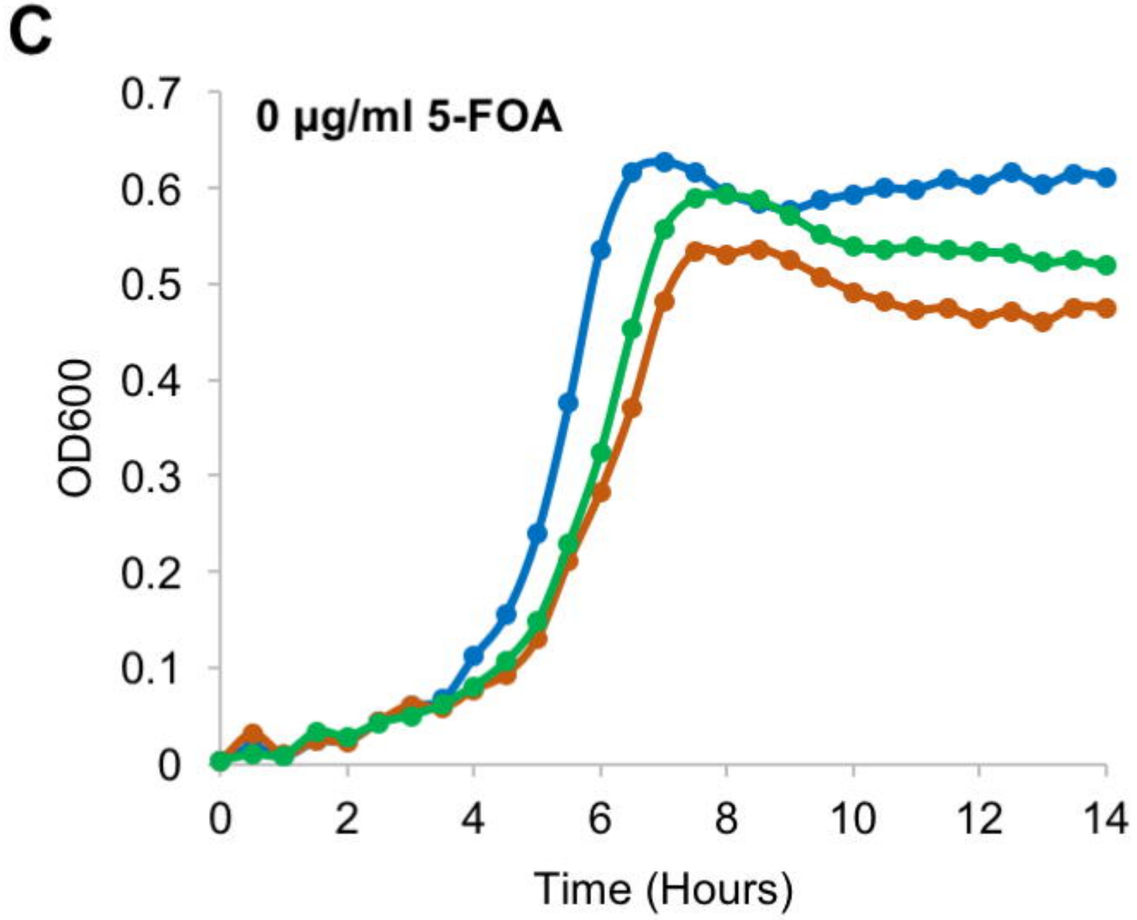


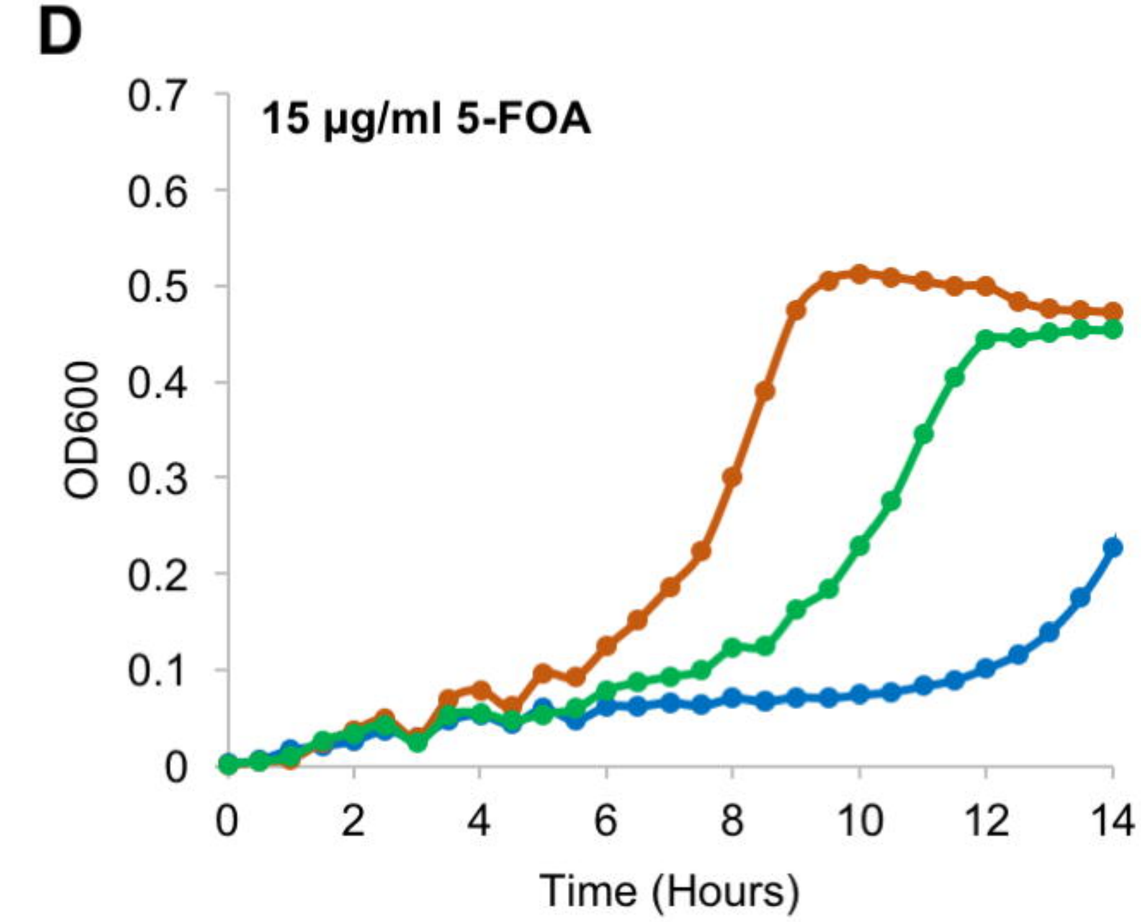

A
PRPP
PyrR
Antitermination
Low [UMP]
High [PRPP]
5'
5'
pyr mRNA
UMP
PyrR
Termination
High [UMP]
Low [PRPP]
5'
B
CDM -Uracil
CDM +Uracil
Read Depth
10000
8000
6000
4000
2000
0
TIGR4
1212
1212.5
1213
1213.5
1214
Kb
SP1276
SP1277
SP1278
C
Fold change
1.2
1
0.8
0.6
0.4
0.2
0
CDM -Uracil
CDM +Uracil

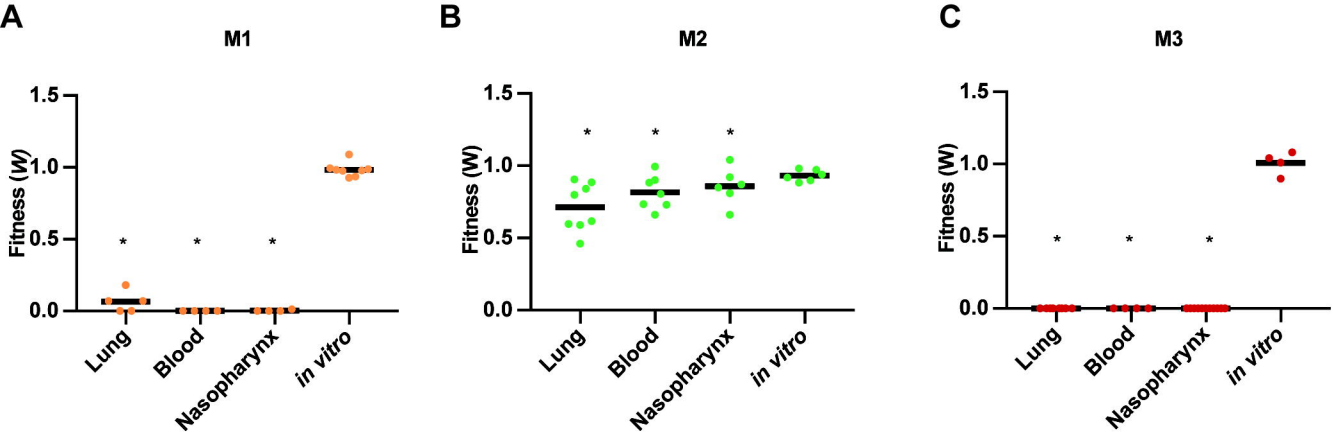

A
M1
bases
deleted
M3
bases
deleted
M2
5'-UAU·GCGCGAUAAAG·UAAAUUUCA-UUUCUUUU-3'
M3
bases
deleted
B
0 µg/ml Uracil
40 µg/ml Uracil
Expression
1.0
0.8
0.6
0.4
0.2
0.0
WTc
M1
M2
M3










A
M1
Fitness (W)
1.5
1.0
0.5
0.0
*
*
*
Lung
Blood
Nasopharynx
in vitro
B
M2
Fitness (W)
1.5
1.0
0.5
0.0
*
*
*
Lung
Blood
Nasopharynx
in vitro
C
M3
Fitness (W)
1.5
1.0
0.5
0.0
*
*
*
Lung
Blood
Nasopharynx
in vitro