1   **The Transcriptional landscape of *Streptococcus pneumoniae* TIGR4 reveals a complex**

2   **operon architecture and abundant riboregulation critical for growth and virulence**

3

4   **Indu Warrier**[1] †**, Nikhil Ram-Mohan**[1] †**, Zeyu Zhu**[1] **, Ariana Hazery**[1] **, Haley Echlin**[2]
5   **Jason Rosch**[2] **, Michelle M Meyer**[1] *****, Tim van Opijnen**[1] *****

6

7   **Running Title**
8   The transcriptional landscape of *Streptococcus pneumoniae*
9

10

11   [1] Biology Department, Boston College, Chestnut Hill, MA, USA
12   [2] Department of Infectious Diseases, St Jude Children's Research Hospital
13
14
15
16
17   † Equal contribution
18   * Corresponding author
19
20   E-mail Addresses:
21
22   MMM: m.meyer@bc.edu
23   TvO: vanopijn@bc.edu

24
25   Keywords:
26   RNA-Seq
27   term-seq
28   5'end-seq
29   Riboregulation
30   sRNA
31   transcriptional landscape
32   operon architecture
33   Streptococcus pneumoniae
34
35

36

37

Abstract

Efficient and highly organized regulation of transcription is fundamental to an organism's ability to survive, proliferate, and quickly respond to its environment. Therefore, precise mapping of transcriptional units and understanding their regulation is crucial to determining how pathogenic bacteria cause disease and how they may be inhibited. In this study, we map the transcriptional landscape of the bacterial pathogen *Streptococcus pneumoniae* TIGR4 by applying a combination of high-throughput RNA-sequencing techniques. We successfully map 1864 high confidence transcription termination sites (TTSs), 790 high confidence transcription start sites (TSSs) (742 primary, and 48 secondary), and 1360 low confidence TSSs (74 secondary and 1286 primary) to yield a total of 2150 TSSs. Furthermore, our study reveals a complex transcriptome wherein environment-respondent alternate transcriptional units are observed within operons stemming from internal TSSs and TTSs. Additionally, we identify many putative *cis*-regulatory RNA elements and riboswitches within 5'-untranslated regions (5'-UTR). 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 demonstrate 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.

Author summary

The canonical relationship between a bacterial operon and the mRNA transcript produced from the operon has become significantly more complex as numerous regulatory mechanisms that impact the stability, translational efficiency, and early termination rates for mRNA transcripts have been described. With the rise of antibiotic resistance, these mechanisms offer new potential targets for antibiotic development. In this study we used a combination of high-throughput sequencing technologies to assess genome-wide transcription start and stop sites, as well as determine condition specific global transcription patterns in the human pathogen *Streptococcus pneumoniae*. We find that the majority of multi-gene operons have alternative start and stop sites enabling condition specific regulation of genes within the same operon. Furthermore, we identified many putative RNA regulators that are widespread in the *S. pneumoniae* pan-genome. Finally, we show that separately collected RNA-Seq data enables identification of conditional triggers for regulatory RNAs, and experimentally demonstrate that our approach may be used to identify drug-able RNA targets by establishing that pyrR RNA functionality is critical for successful *S. pneumoniae* mouse colonization and infection. Thus, our study not only uses genome-wide high-throughput approaches to identify putative RNA regulators, but also establishes the importance of such regulators in *S. pneumoniae* virulence.

79 **Introduction**

80 The transcriptional architecture of bacterial genomes is far more complex than originally

81 proposed. The classical model of an operon describes a group of genes under the control of a

82 regulatory protein where transcription results in a polycistronic mRNA with a single

83 transcription start site (TSS) and a single transcription terminator site (TTS) [1]. However, many

84 individual examples have established that the same operon may encode alternative

85 transcriptional units under varying environmental conditions [2,3]. Furthermore, advancements

86 in sequencing technology that enable highly accurate mapping of TSSs and TTSs on a genome-

87 wide level have demonstrated that the number of TSSs and TTSs can significantly exceed the

88 number of operons [4]. Thus it seems likely that the bacterial transcriptional landscape, or the

89 genome-wide map of all possible transcriptional units, is shaped by an operon architecture that

90 encodes many TSSs and TTSs within single operons, thus significantly increasing complexity

91 with the objective of enabling diverse transcriptional outcomes [5,6].

92

93 To achieve a complex landscape of alternative transcriptional units, transcriptional regulation

94 occurs on multiple levels. In addition to the many protein activators and repressors that control

95 transcription initiation, there are also many non-coding RNAs (ncRNAs), including both small

96 ncRNAs (sRNAs) and highly structured portions of mRNAs that play essential roles as

97 regulatory elements controlling metabolism, stress-responses, and virulence [7-9]. *Trans* – acting

98 small RNAs (sRNAs), which are found in intergenic regions [10] or derived from 3' UTRs [11],

99 allow selective degradation or translation of specific mRNAs [10]. *Cis* – acting mRNA

100 structures, such as riboswitches, which interact with small molecules including metal ions, and

101 protein ligands, and other regulatory sequences that are found in the long 3' UTR of mRNAs,

4

102 affect expression of their respective genes by regulating transcription attenuation or translation

103 inhibition [12,13]. RNA regulation has been shown to play a key role in shaping the

104 transcriptional landscape of a wide range of pathogenic bacteria including *Staphylococcus*

105 *aureus*, *Listeria monocytogenes, Helicobacter pylori,* and strains of *Streptococci* [6,14-22].

106 Several RNA regulators have been validated and associated with pathogenicity and virulence

107 [23,24], and could be used as highly specific druggable targets [25,26], however, only a select set

108 of regulators have been targeted to date [27,28].

109

110 *Streptococcus pneumoniae* is a major causative agent of otitis media, meningitis, pneumonia, and

111 bacteremia. It causes 1.2 million cases of drug-resistant infection in the US annually and results

112 in ~1 million deaths per year worldwide [29-31]. While high-resolution transcriptional mapping

113 data are available for other *Streptococcus* species, these studies have shown limited experimental

114 validation [20], or have focused primarily on the role of sRNAs in virulence [32]. Additionally,

115 previous studies of the *S. pneumoniae* transcriptome have demonstrated the presence of ncRNA

116 regulators and assessed their roles in infection and competence, however, these studies also

117 largely focused on sRNAs [16,18,33]. Thus high-resolution validated maps of the genome wide

118 transcriptional landscape for different *S. pneumoniae* strains would be incredibly valuable.

119

120 Here a comprehensive characterization of the *S. pneumoniae* TIGR4 transcriptional landscape is

121 created using single and paired-end RNA-Seq [34], 5' end-Seq [22], and term-seq (3'-end

122 sequencing) [35] during early-log and mid-log growth in the presence and absence of the

123 antibiotic vancomycin. We obtain a global transcript coverage map, identifying all TSSs, and all

124 TTSs, which highlights a highly complex *S. pneumoniae* transcriptional landscape including

125   many operons with multiple TSSs and TTSs. Furthermore, we demonstrate how TSS and TTS

126   mapping under one set of conditions can be leveraged to analyze independently obtained RNA-

127   Seq data collected under a variety of conditions, and we experimentally validate this approach

128   with several cis-acting RNA regulators. Finally, we demonstrate that the functionality provided

129   by the RNA *cis – regulator pyrR* is critical for *S. pneumoniae in vivo* in a mouse infection model.

130   Importantly, our work demonstrates how a variety of high-throughput sequencing efforts can be

131   combined to map out a comprehensive transcriptional landscape for a bacterial pathogen as well

132   as identify potentially druggable ncRNA targets.

133

134   **Results and Discussion**

135   *Streptococcus pneumoniae has a complex transcriptional landscape.*

136   To characterize the transcriptional landscape of *Streptococcus pneumoniae* TIGR4 (T4), we first

137   determined transcript boundaries by mapping transcription start (TSSs) and termination sites

138   (TTSs) from 5' and 3' end sequencing reads obtained from 3 different conditions: early-log and

139   mid-log growth phase in the absence and presence of vancomycin (Fig 1A and 1B). Since the

140   TSS predictions (both whether a TSS exists and the position identified) correspond well among

141   all three conditions for the vast majority of genes (S1 Fig), reads from both mid-log phase

142   growth conditions were pooled for downstream analyses. To describe a comprehensive

143   transcriptional landscape for T4, the 2341 annotated genes were supplemented by annotating the

144   homologs of known structured non-coding RNAs (ncRNAs) in T4 identified using Infernal (an

145   RNA specific homology search tool [36]), and previously described small RNAs (sRNAs)

146   [16,18] resulting in a total of 2557 features in T4. Given this feature set, a total of 742 primary,

147   and 48 secondary, high confidence TSSs were identified, with a processed coverage greater than

148   100 and a processed/unprocessed ratio greater than 4, where the primary TSS has the highest

149   processed/unprocessed ratio. A total of 1286 low confidence primary TSSs were called that have

150   coverage greater than 2 and a processed/unprocessed ratio greater than 1. In addition, 74 low

151   confidence secondary TSSs that have coverage greater than 100, but a lower

152   processed/unprocessed ratio than the primary TSS were called to yield 2150 total TSSs. From the

153   pooled 3'-end sequencing reads, 1864 TTSs were identified that have a minimum coverage of

154   10, and are enriched at least two-fold over the background (Fig 2, S1 and S2 Tables). Of the 742

155   high confidence TSSs, 625 showed the presence of the GRTATAAT motif at the -10 position

156   and 215 had the MTTGAMAA motif at the -35 position (S1 Fig).

157   In order to assess how the high confidence TSSs and TTSs contribute to transcriptional

158   complexity, we reconstructed T4 transcripts from paired-end RNA-Seq data. StringTie [37,38]

159   detected 343 single gene operons and grouped 1857 genes into 388 multi-gene operons (S3

160   Table, and Fig 3). Based on the reconstructed transcripts, we classified operons into five

161   categories based on the number of internal primary high confidence TSSs and TTSs (Fig 2).

162   Simple operons, transcriptional units with a single gene TSS, and TTS are predominantly

163   identified in the *S. pneumoniae* genome at 47% of all operons (Fig 3A). Traditional operons with

164   multiple genes and a single TSS and TTS make up 10% of the operons, while multiTSS operons

165   make up 1%, and multiTTS operons make up 16%, of all transcriptional units respectively.

166   However, complex operons (most of which consist of two genes) with multiple TSSs and TTSs

167   are the second largest category comprising 26% of all operons (Fig 3A and Fig 3B). Most

168   complex operons are defined by a secondary internal TSS and TTS, however there are several

169   significantly more complex examples where the operon contains multiple TSSs and TTSs (Fig

170   3B), indicating an intricate system of possible transcripts. Of note, transcript prediction from

171   paired-end RNA-Seq using StringTie agrees partially with operon predictions run using only

172   single-end RNA-Seq data [39], or established operon databases that use genome context [40].

173   Such discrepancies between strategies for operon calling have been observed previously (e.g.

174   *Escherichia coli* MG1655) [4,41]). For example, StringTie groups genes SP_0001-SP_0014 into

175   a single transcript while Rockhopper splits them into three operons. However, it is evident from

176   the paired-end RNA-Seq coverage map (see S2 Fig) that despite the fluctuations in coverage

177   between the coding regions, there is still large enough expression in the intergenic regions for

178   StringTie to group these genes together. This example demonstrates the role of internal

179   transcriptional features in creating alternative transcriptional units. Of note, operon predictions

180   based on either method can only be validated by directed, low-throughput approaches. However,

181   despite the differences arising from the methodologies used, the complexity in operon structure

182   we deduce remains nearly the same except for the portion of multiTTS operons (see S2 Fig). The

183   TSS, TTS, our supplemented annotations, and reconstructed transcripts can be easily visualized

184   using files available at https://github.com/nikhilram/T4pipeline.

185

186   Since our data revealed many operons with complex structure, we sought to corroborate specific

187   examples using additional data sources. One complex operon we identified, which is also present

188   in existing databases of operon structure [5,42], consists of 9 genes (SP_1018-SP_1026)

189   encoding thymidine kinase, GNAT family N-acetyltransferase, peptide chain release factor 1,

190   peptide chain release factor N(5)-glutamine methyltransferase, threonylcarbomyl-AMP synthase,

191   N-acetyltransferase, serine hydroxymethyltransferase, nucleoid-associated protein, and

192   Pneumococcal vaccine antigen A respectively, with two primary high confidence and six other

193   internal TSSs and eight TTSs (Fig 4A). In addition, this operon displays unequal and complex

194　　gene expression patterns when independently collected RNA-Seq data from diverse media

195　　conditions is mapped to the transcript. In poor growth medium (MCDM) the operon can be split

196　　into two parts based on expression observed from RNA-Seq coverage maps, where the last five

197　　genes in the operon (SP_1022-1026) are expressed higher than the first four genes (SP_1018-

198　　1021), while in rich medium (SDMM) the read depth across the operon is similar. While none of

199　　the individual genes show significant differential expression between the two conditions (>2 fold

200　　change [43,44], there is a drastic difference between the two conditions in the average coverage

201　　per gene amongst the first 4 versus the last 5 genes. In rich medium, SP_1022-1026 show an

202　　average coverage of ~1.9x more than SP_1018-1021, while in the poor medium the average

203　　coverage of SP_1022-1026 is ~3.6x that of SP_1018-1021. This observation corroborates the

204　　role of the internal regulatory mechanisms to create alternate transcriptional units for maintaining

205　　differences in gene expression between different growth conditions especially since we identify a

206　　high confidence TSS upstream of SP_1022.

207

208　　Additional validation of our data and analysis approach derives largely from existing low-

209　　throughput experiments. The *mal* regulon is a multiple operon system under the control of a

210　　single protein, *malR* (SP_2112), which downregulates regulon expression at the *malM*

211　　(SP_2107) promoter [45]. Our data shows that the mal regulon includes operons belonging to

212　　two different categories, a multiTTS operon (*malMP*/SP_2016-2107) and a complex operon

213　　(*malXCD*/SP_2108-2112). From the RNA-Seq coverage maps it is clear that the two operons can

214　　be differentially controlled and expressed in rich vs poor media (Fig 4B). Again, although the

215　　genes are not significantly differentially expressed in the two conditions [43,44], the average

216　　coverage for each gene is greater in the poor medium in comparison to that in the rich medium.

217 Furthermore, the TSS and TTS identified by our analysis reveal features that have been

218 previously described in lower throughput assays [24]. Despite experimental analyses grouping

219 SP_2108-2110 and SP_2111-2112 into two different operons, it is evident from the paired-end

220 RNA-Seq coverage map as well as RNA-Seq of T4 grown in MCDM that transcription continues

221 past the TTS of SP_2110 into SP_2111. In SDMM, however, transcription seems to mimic what

222 has been experimentally shown. Thus, although our data may highlight many examples of

223 complex transcriptional architecture, these examples are verifiable through the incorporation of

224 additional RNA-Seq data, and where applicable are consistent with low-throughput studies done

225 in the past.

226

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

228 To identify RNA regulators that act through premature transcription termination, we compiled 3'

229 end sequencing reads upstream of translational start sites, allowing a minimum 5'-untranslated

230 region (UTR) length of 70 bases. We detected 162 such early TTS sites that represent novel high

231 and low confidence putative regulatory elements, of which 141 were completely intergenic

232 (represented in blue in the 5th band of Fig 2). Intriguingly, these putative regulatory elements are

233 not limited to the 5'-UTR of protein coding genes. Of these 141 novel putative regulatory

234 elements, 129 are upstream of protein coding genes while the rest are in the 5'-UTR of annotated

235 ncRNAs. A ncRNA under the regulation of another ncRNA has been shown in *Enterococcus*

236 *faecalis* [46] and *Listeria monocytogenes* [47] but this is the first putative example of this

237 mechanism in *Streptococcus pneumoniae* TIGR4. By screening these regulatory elements against

238 380 published *S. pneumoniae* strains [48], covering a large part of the pan-genome, we found

239 that 80 candidates (~57%) were conserved across all genomes, 107 (76%) were identified in at

240 least 350 genomes, 12 (~9%) candidates were identified in fewer than half of the genomes (S3A

241 Fig), while the least distributed candidate, identified upstream of a hypothetical protein, was

242 found only in 52 genomes analyzed. Interestingly, 140 (99%) candidates were found as single

243 copies within a genome, while copies of the putative candidate upstream of the large subunit of

244 the rRNA were found preceding each of the 3 rRNA copies in each genome. Evolutionary

245 distance of each candidate cluster was estimated using MEGA-CC [49], which reveals that each

246 cluster is made of highly similar, if not identical, sequences (S3B Fig).

247

248 *Leveraging RNA-Seq data collected under various conditions enables identification and*

249 *validation of environment-responsive RNA regulators.*

250 We reasoned that since we mapped RNA-regulators by means of 5'end and 3'end sequencing,

251 we would be able to associate these regulators with specific growth conditions using

252 environment dependent RNA-Seq data. To confirm the biological relevance of an RNA regulator

253 and associate it with a specific condition one would expect to see a change in the 5' UTR

254 coverage relative to the accompanying gene. For instance, if a regulator forms an early

255 terminator the RNA-Seq coverage in the 5'UTR is relatively high, while the coverage in the

256 controlled gene would be much lower. Alternatively, if the environment relieves the formation of

257 the early terminator the coverage across the 5'UTR and gene would become less skewed. To

258 determine the applicability of this assumption we leveraged independently collected RNA-Seq

259 data sampled under different nutrient conditions including, rich and poor media [44], and

260 nutrient depletion conditions where a single nutrient was removed from the environment. RNA-

261 Seq data were mapped to each putative regulatory region and coverage was calculated and

262 averaged across the length of the 5' UTR regulatory element and the downstream gene. From our

263   list of candidate 5'-UTR regulatory elements, 36 showed more than two fold change in read-

264   through between rich and poor media, with the majority showing an increase in readthrough in

265   poor media. Furthermore, as more RNA-Seq data from other conditions become available, this

266   list is likely to grow to include more putative RNA regulators that show differential read-through

267   under alternative conditions.

268

269   To associate a regulator to a specific growth condition, we validated homologs of two previously

270   identified RNA *cis*-regulators; the thiamine pyrophosphate (TPP) (preceding SP_0716) and flavin

271   mononucleotide (FMN) (preceding SP_0178) riboswitches, using qRT-PCR (Fig 5A, C),

272   confirming their media-specific changes in transcription. In many bacteria the TPP riboswitch

273   binds thiamine pyrophosphate and regulates thiamine biosynthesis and transport [50]. Similarly,

274   the FMN riboswitch regulates biosynthesis and transport of riboflavin by binding to FMN [51].

275   While we validated that these riboswitches respond to poor media by increasing expression of

276   their respective genes (Fig 5A, C) we suspected that this was due to depletion of each specific

277   ligand in the poor media. Indeed, when poor media is supplemented either with thiamine or

278   riboflavin, expression of the TPP or FMN controlled gene (SP_0716 and SP_0178 respectively)

279   decreases by more than 3-fold (Fig 5 B, D), suggesting that the observed differences between

280   rich and poor media can be directly attributed to the activity of these riboswitches.

281

282   In addition to previously well-known RNA regulators, a putative regulator preceding ribosomal

283   protein L11 methyltransferase (SP1782) that showed a decrease in read-through and expression

284   in poor media was validated (Fig 5E). While the magnitude of the expression change apparent

285   from qPCR is not as strong as that calculated from the RNA-Seq coverage, there are significant

286   changes in expression that matches the direction of change. Since many ribosomal proteins are

12

287   regulated through the action of cis-regulatory RNAs [52], this seems like a strong candidate for

288   further study. However, due to the considerable challenges associated with identifying the exact

289   ligand for this putative regulator, a definitive biological function has yet to be assigned and is

290   currently under investigation.

291

292   In an attempt to validate the feasibility of directly associating RNA-regulators with a highly

293   specific change in the environment we performed RNA-Seq in the presence and absence of

294   uracil. One specific regulatory element that is sensitive to uracil is the pyrR RNA element, which

295   in many bacteria regulates *de novo* pyrimidine nucleotide biosynthesis through a transcription

296   attenuation mechanism mediated by the PyrR regulatory protein [53,54]. In the presence of the

297   co-regulator UMP, PyrR binds to the 5' UTR of the *pyr* mRNA transcript (the pyrR RNA

298   element) and disrupts the anti-terminator stem-loop thereby promoting the formation of a factor-

299   independent transcription terminator resulting in reduced expression of downstream genes [53]

300   (Fig 6A). In contrast, the co-regulator 5-phosphoribosyl-1-pyrophosphate (PRPP) antagonizes

301   the action of UMP on termination by binding to the PyrR protein when UMP concentration is

302   low [55] (Fig 6A). For *S. pneumoniae*, our data confirms that pyrR RNA elements are present in

303   the 5' UTR of two *pyr* operons (SP_1278-1276; SP_0701-0702), and the uracil transporter

304   (SP_1286). Furthermore, in response to the absence of uracil the coverage across the two genes

305   directly adjacent to the regulators (SP_1278 and SP_0701) and over the entire two operons

306   increases drastically (Fig 6B), demonstrating that the regulatory elements effectively turn the

307   genes/operons on, which we confirmed by qRT-PCR (Fig 6C). Thus, while term-seq can be used

308   to map novel regulatory RNA candidates on a genome-wide scale, RNA-Seq data can be

309   leveraged, even in retrospect, to identify environmental conditions the regulator responds to.

310

311   *The pyr operon is regulated through the secondary structure of the 5' RNA leader-region, is*

312   *essential for in vitro growth and in vivo virulence and can be directly manipulated.*

313   To further investigate the importance of the pyrR regulatory RNA element in growth, three

314   different mutants were constructed that variably affect the 5' RNA secondary structure (Fig 6D):

315   1) mutation M1 interferes with the binding of PyrR to the *pyr* mRNA; 2) mutation M2 renders

316   the regulatory element in an "always on" state by destabilizing the rho-independent terminator

317   stem-loop structure that is formed in the presence of UMP; 3) M3 locks the terminator and

318   creates an "always off" state (Fig 6D). Wild type and mutant strains were cultured in the

319   presence or absence of uracil and the effect of the mutations on expression of SP_1278 were

320   assessed with qRT-PCR (Fig 6E) and further confirmed by β-galactosidase reporter assays (S4

321   Fig). As expected, expression in the wild type decreased (9.5-fold) in the presence of uracil

322   confirming the repressive effect of exogenous pyrimidine (Fig 6E) [56]. M1, which should be

323   insensitive to the presence of PyrR and its co-regulator UMP (Fig 6D) is indeed unresponsive to

324   the presence of uracil (Fig 6E). M2 triggers constitutive expression of the *pyr* operon (Fig 6E)

325   and M3 has a ~5-fold reduction in expression compared to the wild type regardless of the

326   presence of uracil (Fig 6E).

327

328   Previously we showed that the pyrimidine synthesis pathway in *S. pneumoniae* is partially

329   regulated by a two-component system (SP_2192-2193) and that genes in this pathway are

330   important for growth [57]. To determine the importance of a functional pyrR regulatory RNA

331   element in growth, we performed growth experiments with mutants M1, M2 and M3 in the

332   absence and presence of uracil. These data suggest that a functional pyrR does not appear to be

14

333   absolutely necessary. For instance, while M1 may have a slight growth defect when cultured in

334   the absence of uracil, M2 has no growth defect in the presence or absence of uracil (Fig 7A, B).

335   Although both mutations result in constitutive expression of the *pyr* operon, mutation M1 leads

336   to higher expression (Fig 6E, S4 Fig) indicating that overexpression of the pyr genes may result

337   in accumulation of end products that are detrimental to the cell. Alternatively, the M2 pyrR RNA

338   element can still bind excess UMP-bound PyrR (as its PyrR binding domain is intact) thus

339   reducing the effective concentration of UMP in the cell and thereby potential accumulation

340   associated side-effects. Importantly, M3 has a severe growth defect compared to wild type in the

341   absence of uracil (Fig 7A) ($p<0.002$), which can be partially rescued upon addition of uracil (Fig

342   7B). This suggests that while a constitutive off-state is detrimental for the bacterium in the

343   absence of uracil a constitutive on-state can be overcome, indicating that efficient transcriptional

344   control may not be essential.

345

346   To determine whether we can manipulate the manner in which the pyrR RNA element affects

347   growth, we determined growth in the presence of 5-Fluoroorotic acid (5-FOA), a pyrimidine

348   analog. 5-FOA is converted into 5-Fluorouracil (5-FU) a potent inhibitor of thymidylate

349   synthetase, whose activity is essential for DNA replication and repair [58]. Additionally, 5-FU

350   competes with UMP for interacting with the PyrR protein [59]. 5-FU can thus work as a decoy,

351   signaling that UMP is present in the cell; triggering the formation of a terminator and reducing

352   expression of the *pyr* operon. The wild type strain displayed a severe growth defect in the

353   presence of 5-FOA (Fig 7C&D) ($p<0.05$), while M1 (which should not interact with PyrR and

354   should thus be largely insensitive to the presence of 5-FOA) displayed a much smaller growth

355   defect (Fig 7C&D). In addition, M2, which constitutively over expresses the *pyr* operon, is also

356   less sensitive to 5-FOA then wild type (Fig 7C&D) (*p<0.05*). Thus, the mutations we introduced

357   into the pyrR RNA element affect the secondary structure in the manner that we intended, and

358   can have far reaching regulatory and fitness effects. Importantly, it shows that a drug targeted

359   against the secondary structure can directly manipulate and severely hamper growth.

360

361   A remaining key question is the importance of RNA regulatory elements in colonization and the

362   induction of disease. Somewhat surprisingly our *in vitro* growth curves suggest that constitutive

363   expression of the *pyr* operon (M1) and constitutive overexpression (M2) is not substantially

364   detrimental to growth, indicating that efficient regulation is not critical. To assess the effect of

365   loss of regulation on bacterial fitness *in vivo*, the pyrR RNA element mutants were tested in 1x1

366   competition assays (mutant vs. wild type) in a mouse infection model (Fig 8). While fitness for

367   all three mutants is similar to the unmodified strain *in vitro* in the presence of uracil, M1 and M3

368   are unable to colonize and survive in the mouse nasopharynx, or infect and survive in the lung

369   and transition and survive in the blood (Fig 8A&C). M2 has less of a defect *in vivo*, but still has a

370   significantly diminished ability to infect and survive in the lung (Fig 8B). These results indicate

371   that efficient regulation of the *pyr* operon *in vivo* is critical for growth and survival of *S.*

372   *pneumoniae* within the host. While we had previously shown that genes in the *pyr* operon are

373   important *in vivo* [57]*,* the regulatory findings in this project take our understanding a step

374   further and, importantly, in combination with the findings that 5-FOA can efficiently interact

375   with the RNA regulatory element, suggests that it is feasible to modulate *in vivo* fitness and

376   thereby virulence by targeting such regulatory elements.

377

16

378 **Towards comprehensive transcriptional landscape reconstructions and highly targeted**

379 **regulatory RNA element inhibitors.**

380 With the advent of deep sequencing technologies, our understanding of prokaryotic

381 transcriptional dynamics is rapidly advancing [60] and underlining that bacterial transcriptomes

382 are not as simple as previously thought. Analysis of the *S. pneumoniae* TIGR4 transcriptome

383 using three different sequencing techniques (RNA-Seq, term-seq, and 5'end-Seq) has led to a

384 comprehensive mapping of its transcriptional landscape. We successfully map 742 primary and

385 48 secondary high confidence TSSs and 74 low confidence secondary TSSs (2150 total) and

386 1864 TTSs. These results along with a recent study that annotates the *S. pneumoniae* D39

387 genome [61], deeply enriches the understanding of transcriptional regulation in this important

388 pathogen. Besides identifying the transcript boundaries, we also uncovered a complex operon

389 structure, which has also been found in *E. coli* [4]. Importantly, such complexity likely allows

390 for environment-dependent modulation of gene expression producing variable transcripts in

391 response to varying conditions, which we illustrated here through analyses of a 9-gene complex

392 operon and the mal regulon (Fig 4). Additionally, similar environment-dependent versatile

393 operon behavior has been observed in *E. coli* [4] and to a lesser extent in *Mycoplasma*

394 *pneumoniae* [62]. This means that our understanding is shifting dramatically and it is thus

395 becoming clear that operons in bacteria should be seen as adaptable structures that can

396 significantly increase the regulatory capacity of the transcriptome by responding to

397 environmental changes in a highly specific manner. Furthermore, this technology can potentially

398 be extended to *in vivo* growth environments to determine how the bacteria modulates its gene

399 expression in the context of the host immune system. A major challenge with respect to such

400 procedures is the quantity and quality of RNA to generate TSSs, TTSs and RNA-Seq coverage

17

401   data. However, such limitations could be overcome by obtaining only RNA-Seq coverage data

402   and leveraging these data, as we have done in this study, to identify RNA regulators specific to

403   *in vivo* environments.

404

405   Another central aspect of our approach is the identification of putative 5'-UTR structured

406   regulatory elements. Riboswitches and other untranslated regulatory elements (binding sites for

407   small regulatory RNAs) are important bacterial RNA elements that are thought to regulate up to

408   2% of bacterial genes [7,8]. However, the discovery of new regulators is difficult when relying

409   solely on computational methodology and sequence conservation [63]. Here we show that

410   through term-seq [35] it is possible to identify such RNA elements on a genome-wide scale and

411   by combining it with RNA-Seq performed in different conditions transcriptional phenotypes can

412   be directly linked to the RNA element. This strategy thus makes it possible to screen for

413   regulatory RNA elements in retrospect by making use of already existing or newly generated

414   RNA-Seq data.

415

416   Importantly, besides the ability to re-construct an organism's intricate transcriptional landscape

417   we show that there is also a direct application of our multi-sequencing approach, namely the

418   ability to inhibit operons and/or pathways with specific chemicals or drugs that target the RNA

419   regulatory element. We show that this is possible for the pyrR RNA element, a regulatory

420   element that is important for pneumococcal growth and virulence, which means that this

421   regulatory element could be a potential antimicrobial drug target. This idea is further

422   strengthened by the fact that *S. pneumoniae* displays a growth defect in the presence of 5-FOA,

423   which directly relates to misregulation of pyrR RNA confirming its drug-able potential.

424

425   We believe that the presented multi-omics sequencing strategy brings a global understanding of

426   regulation in *S. pneumoniae* significantly closer, and because the approach is easily transferable

427   to other species, it will enable species-wide comparisons for conservation of operon structure and

428   regulatory elements. In addition, such detailed regulatory understanding creates new regulatory

429   control tools for synthetic biology purposes. Moreover, the combination with *in vivo* experiments

430   shows that it is a realistic goal to design or select specific compounds that target ribo-regulators

431   in order to mitigate virulence or antibiotic resistance.

432

433   **Methods:**

434   *Culture conditions and sample collection*

435   For RNA-Seq, term-seq and 5'end-Seq library preparation, *Streptococcus pneumoniae* TIGR4

436   (T4) was cultured in rich media (SDMM) to mid-log phase ($OD_{600} = 0.4$). Cultures were diluted

437   to an OD_{600} of 0.05 in fresh media, grown for one doubling (T_0). Cultures were then grown in the

438   presence or absence of 0.24 µg/ml vancomycin. At 0 min (T_0) and after 30 min of growth (T_{30})

439   10ml culture was harvested by means of centrifugation (4000 rpm, 7 min at $4^{o}C$) followed by

440   flash freezing in a dry-ice ethanol bath and storage at $-80^{o}C$ until RNA extraction. Sample

441   collection was performed in four biological replicates and total RNA was isolated using an

442   RNeasy Mini kit (Qiagen). For qRT-PCR analyses, T4 was cultured in SDMM to mid-log phase

443   ($OD_{600} = 0.4$) and after centrifugation cultures were washed with 1X PBS and diluted to an

444   OD_{600} of 0.003 in appropriate media. Cultures were harvested at mid-log followed by RNA

445   extraction as described above.

446

447   *5'end-Seq library preparation*

448   5'end-Seq libraries were generated by dividing the total RNA into 5' polyphosphate treated

449   (processed) and untreated (unprocessed) samples that were subsequently processed and

450   sequenced according to protocols described in [22] and [64] with few modifications. See

451   supplemental methods for a detailed protocol.

452

453   *RNA-Seq library preparation*

454   RNA-Seq libraries were generated by using the RNAtag-Seq protocol [34,44]. Briefly, 400 ng

455   RNA was fragmented in FastAP buffer, DNase-treated with Turbo DNase, 5'-dephosphorylated

456   using FastAP. Barcoded RNA adapters were then ligated to the 3' terminus, samples from

457   different conditions were pooled and ribosomal RNA was depleted using the Ribo-zero rRNA

458   removal kit. Illumina cDNA sequencing libraries were generated by first-strand cDNA synthesis,

459   3' linker ligation and PCR with Illumina index primers for 17 cycles. The final concentration and

460   size distribution were determined with the Qubit dsDNA BR Assay kit and the dsDNA D1000

461   Tapestation kit, respectively.

462

463   *Paired-end RNA-seq library preparation*

464   RNA was extracted in triplicate from cells grown to mid-logarithmic phase (OD_{620}=0.4) using

465   the QiaShredder and RNeasy Kit (Qiagen) by following the kit instructions with the exception of

466   the lysis step, which was performed by membrane disruption using zirconia/silica beads. After

467   extraction, RNA was treated with DNA Free DNA removal kit (Thermo) to remove DNA. RNA-

468   seq library was generated via ScriptSeq vs RNA-seq library kit (epicenter) using 1 ug RNA. The

469   kit instructions were followed using Ampure beads (Agencourt) for cDNA and library

470   purification. Each replicate was individually indexed using Failsafe PCR enzyme (Epicentre).

471   Completed libraries were analyzed for insert size distribution on a 2100 BioAnalyzer High

472   Sensitivity kit (Agilent Technologies, Santa Clara, California). Libraries were quantified using

473   the Quant-iT PicoGreen ds DNA assay (Life Technologies,). One hundred cycle paired end

474   sequencing was performed on an Illumina HiSeq 4000.

475

476   *term-seq library preparation*

477   term-seq libraries were generated as previously described [35] with few modifications. 2 µg total

478   RNA was depleted of genomic DNA using Turbo DNase, 5' dephosphorylated, ligated to

479   barcoded RNA adapters at the 3' terminus and fragmented in fragmentation buffer. Barcoded and

480   fragmented RNA from different conditions were pooled and ribosomal RNA was depleted using

481   Ribo-zero. cDNA libraries were generated by first strand cDNA synthesis and RNA template

482   was degraded as mentioned in the 5'end-Seq library preparation. Second 3' linker was ligated

483   and PCR amplified with Illumina index primers for 17 cycles. All four library preparations

484   (RNAtag-Seq, term-seq, 5'end-Seq processed and 5'end-Seq unprocessed) were pooled

485   according to the method of preparation and sequenced at high depth (8.5 million reads/sample)

486   on an Illumina NextSeq500.

487

488   *Read processing and mapping*

489   The *S. pneumoniae* TIGR4 (NC_003028.3) genome annotation was supplemented with all

490   known structured ncRNAs and sRNAs. Homologs of known ncRNA families deposited in Rfam

491   [65,66] were identified in the genome by the cmsearch function of Infernal 1.1 [36]. sRNAs

492   previously identified in *S. pneumoniae* [16,18] were identified in T4 using BLAST [67].

493   Coordinates for the hits from the above searches are included in the supplemented annotation.

494   The single-end sequencing reads from the 5' end-Seq sequencing, 3' end sequencing (term-seq),

495   and RNAtag-Seq were processed and mapped to the supplemented *S. pneumoniae* TIGR4

496   (NC_003028.3) genome using the in-house developed Aerobio pipeline. Aerobio runs the

497   processing and mapping in two phases. Phase 0 employs bcl2fastq to convert BCL to fastq files,

498   quality control and de-multiplexing and compilation of the reads based on the sample conditions.

499   Phase 1 maps the de-multiplexed reads against the genome, under default parameters, using

500   Bowtie2 [68] and streams the uniquely mapped reads to SAMtools [69] to generate sorted and

501   indexed BAM files for each sample.

502

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

504   Perl code from [35] was adapted to map the strand specific single nucleotide coverage of the 5'

505   positions of the 5' end and 3' end sequencing reads respectively estimated using BEDTools [70].

506   With this nucleotide level coverage data calculated from the 5' end-Seq, regions up to 500

507   nucleotides upstream of the translational start sites described in the annotated TIGR4 genome

508   (NC_003028.3) were scanned for mapped reads with a minimum coverage of 2 and a

509   processed/unprocessed ratio of 1 as in [35]. Predicted primary TSSs were called as high

510   confidence if the position had a minimum processed coverage greater than 100 and a

511   processed/unprocessed ratio of 4. Secondary TSSs are greater than 10 nucleotides up or

512   downstream of predicted primary TSSs. When multiple putative TSSs were identified in a 5'

513   UTR, the higher confidence, or the highest processed/unprocessed ratio was assigned as the

514   primary TSS for the downstream gene.

515 Regions up to 100 nucleotides upstream of the predicted high confidence TSSs were split into

516 three datasets – TSS-20; TSS-20-50; and TSS-100, and searched for enriched promoter motifs

517 and transcription factor binding sites listed in [71] using MEME [72]. Similar to the

518 identification of the TSSs, TTSs were identified by scanning up to 150 nucleotides downstream

519 of the translational stop site for mapped 3' end reads with a minimum coverage of 10 in the

520 pooled dataset. The position with the highest coverage was considered the most likely TTS for a

521 gene. Predicted TTSs were called as high confidence if the coverage of the site was at least

522 double the average coverage in the 150 nucleotide region. 40 nucleotides upstream of the

523 predicted TTS were scanned for a stable stem-loop structure using RNAFold [73], and the

524 number of uridines in the 8 nucleotides immediately upstream calculated. The modified perl

525 scripts, sample input files, and the track files for visualization are available at

526 https://github.com/nikhilram/T4pipeline.

527

528 *Identification of transcript boundaries and operon structures in the genome*

529 Paired-end reads were mapped to the T4 genome using TopHat2 [74]. The BAM file of the

530 mapped reads were quality filtered and used to predict transcript boundaries and group genes into

531 operons using a stringent run of StringTie [37,38]. Specifically, 0 locus gap separation was

532 allowed and a minimum junction coverage of 10 was required. Predicted operons were compared

533 with the genome based predictions listed in the Database of Prokaryotic Operons [5,42,75], and

534 complexity in the operon structure was characterized by surveying the number of high

535 confidence primary internal TSSs and enriched TTSs similar to [4].

536

537 *Identification of candidate regulatory elements*

23

538   Once the TSSs were identified, 5'-UTR regions with a length of at least 70 nucleotides were

539   scanned for mapped 3' end sequencing reads with a minimum coverage of 2 to identify putative

540   early terminators. 5'-UTR regions with a predicted early TTS were binned as candidate

541   regulatory elements. The nucleotide sequence for each candidate element was obtained and

542   folded using RNAFold [73]. Secondary structures and free energy values were compiled for each

543   candidate.

544

545   The response of candidate regulatory elements to different media conditions were assessed by

546   calculating the RNA-Seq coverage in both the regulatory element and the regulated gene. Read-

547   through was calculated for each of the candidates as described previously [35]. Briefly, read-

548   through is the ratio (denoted in percentage) of the average coverage across the gene to that of the

549   5'-UTR identified here. The greater the read-through, the higher the expression of the gene with

550   respect to the 5'-UTR. That is, if the regulator reduced the expression of the gene, read-through

551   would be small. If the regulator turned on gene expression in response to certain conditions, the

552   read-through would be large.

553

554   *Conservation of the candidate regulatory elements in Streptococcus pneumoniae*

555   A local BLAST [67] database was generated with the genomes of 30 *S. pneumoniae* strains

556   available in Refseq 77 [76] and 350 strains from [48]. Each of the candidate regulators identified

557   in the genome of TIGR4 was BLASTed against this database, and hits in the other genomes were

558   extracted and aligned using MAFFT version 7 [77]. The degree of conservation across the 380

559   genomes was determined by surveying each candidate cluster post filtering to remove sequences

560   that were less than 70% in length of the query and with e-values greater than 1×10^{-4}.

561

562   *Expression analysis using qPCR*

563   RNA was isolated from cultures using the Qiagen RNeasy kit (Qiagen). DNase treated RNA was

564   used to generate cDNA with iScript reverse transcriptase supermix for RT-qPCR (BioRad).

565   Quantitative PCR was performed using a Bio-Rad MyiQ. Each sample was normalized against

566   the 50S ribosomal gene, SP_2204 and were measured in biological replicate and technical

567   triplicates. No-reverse transcriptase and no-template controls were included for all samples.

568   Primers used in this study are listed in Table S7.

569

570   *pyrR RNA mutant strain construction*

571   To generate a pyrR RNA mutant strain, wild-type pyrR RNA element region and ~ 1 kb regions

572   of homology flanking on either side of the RNA element was PCR amplified from *S.*

573   *pneumoniae* TIGR4 genomic DNA (GenBank accession number NC_003028.3). Mutations to

574   the pyrR RNA element were obtained either by site-directed mutagenesis or by PCR assembly

575   using appropriate primers (S7 Table). A chloramphenicol resistance cassette was PCR amplified

576   from pAC1000. The amplified products were assembled such that the chloramphenicol resistance

577   cassette was inserted immediately upstream of the promoter in the opposite orientation. The

578   assembled products were transformed into *S. pneumoniae* TIGR4 as previously described [78]

579   and transformants were screened for resistance to chloramphenicol and mutation verified by

580   Sanger sequencing.

581

582   *pyrR RNA mutant growth assays*

583   Wild type and pyrR RNA mutants of T4 were grown for 2 hours and diluted to an OD_{600} of 0.015

584   in fresh media, with varying concentrations of uracil and/or 5-FOA. Growth assays were

585   performed in 96-well plates for 16 hours by taking OD600 measurements every half hour using a

586   Tecan Infiniti Pro plate reader (Tecan). Growth assays were performed no less than three times.

587

588   *In vivo pyrR mutant fitness determination*

589   1 x 1 competition experiments were performed with pyrR RNA mutants (M1 to M3) that were

590   competed against the wild-type strain after which bacterial fitness was calculated as previously

591   described [57] with a few modifications. Lung removal and homogenization (in 10 ml 1 x PBS),

592   blood collection (100 µl) and nasopharnyx lavage (with 1 ml 1X PBS) were perfromed on all

593   animals 24 hours post infection, with the exception of pyrR M3, which due to the large fitness

594   defect were harvested at 6 hours post infection. Experiments involving animals were performed

595   in accordance with guidelines of IRB/IACUC and approval of the Boston College Animal

596   Facility.

597

604

605   **References**

606
607   1.   Jacob F, Manod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol.
608   1961;3: 318–356. doi:10.1016/S0022-2836(61)80072-7

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

611 3.   de Saizieu A, Certa U, Warrington J, Gray C, Keck W, Mous J. Bacterial transcript
612   imaging by hybridization of total RNA to oligonucleotide arrays. Nat Biotech. 1998;16:
613   45–48. doi:10.1038/nbt0198-45

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

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

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

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

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

627 9.   Lebreton A, Cossart P. RNA- and protein-mediated control of Listeria monocytogenes
628   virulence gene expression. RNA Biology. 2017;14: 460–470.
629   doi:10.1080/15476286.2016.1189069

630 10. Saberi F, Kamali M, Najafi A, Yazdanparast A, Moghaddam MM. Natural antisense
631   RNAs as mRNA regulatory elements in bacteria: a review on function and applications.
632   Cell Mol Biol Lett. BioMed Central; 2016;21: 6. doi:10.1186/s11658-016-0007-z

633 11. Chao Y, Vogel J. A 3' UTR-Derived Small RNA Provides the Regulatory Noncoding Arm
634   of the Inner Membrane Stress Response. Molecular Cell. 2016;61: 352–363.
635   doi:10.1016/j.molcel.2015.12.023

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

638 13. Ruiz de los Mozos I, Vergara-Irigaray M, Segura V, Villanueva M, Bitarte N, Saramago
639   M, et al. Base pairing interaction between 5"- and 3-"UTRs controls icaR mRNA
640   translation in Staphylococcus aureus. PLoS Genet. 2013;9: e1004001.
641   doi:10.1371/journal.pgen.1004001

642 14. Eyraud A, Tattevin P, Chabelskaya S, Felden B. A small RNA controls a protein regulator
643   involved in antibiotic resistance in *Staphylococcus aureus.* Nucleic Acids Res. 2014;42:
644   4892–4905. doi:10.1093/nar/gku149

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

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

650   17.   Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, et al. CRISPR
651   RNA maturation by trans-encoded small RNA and host factor RNase III. Nature.
652   2011;471: 602–607. doi:10.1038/nature09886

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

693   33.   Kumar R, Shah P, Swiatlo E, Burgess SC, Lawrence ML, Nanduri B. Identification of
694   novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-
695   resolution genome tiling arrays. BMC Genomics. 2010;11: 350. doi:10.1186/1471-2164-
696   11-350

697   34.   Shishkin AA, Giannoukos G, Kucukural A, Ciulla D, Busby M, Surka C, et al.
698   Simultaneous generation of many RNA-seq libraries in a single reaction. Nat Meth.
699   2015;12: 323–325. doi:10.1038/nmeth.3313

700   35.   Dar D, Shamir M, Mellin JR, Koutero M, Stern-Ginossar N, Cossart P, et al. Term-seq
701   reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science. 2016;352:
702   187–187. doi:10.1126/science.aad9822

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

705   37.   Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie
706   enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotech.
707   Nature Publishing Group; 2015;33: 290–295. doi:10.1038/nbt.3122

708   38.   Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis
709   of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. Nature
710   Publishing Group; 2016;11: 1650–1667. doi:10.1038/nprot.2016.095

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

713   40.   Mao X, Ma Q, Zhou C, Chen X, Zhang H, Yang J, et al. DOOR 2.0: presenting operons
714   and their functions through dynamic and integrated views. Nucleic Acids Res. 2014;42:
715   D654–9. doi:10.1093/nar/gkt1048

716   41.   McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, et al.
717   Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res. 2013;41: e140–
718   e140. doi:10.1093/nar/gkt444

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

721   43.   Anders S, Huber W. Differential expression analysis for sequence count data. Genome
722   Biol. BioMed Central; 2010;11: R106. doi:10.1186/gb-2010-11-10-r106

723   44.   Jensen PA, Zhu Z, van Opijnen T. Antibiotics Disrupt Coordination between
724   Transcriptional and Phenotypic Stress Responses in Pathogenic Bacteria. CellReports.
725   2017;20: 1705–1716. doi:10.1016/j.celrep.2017.07.062

726   45.   Nieto C, Espinosa M, Puyet A. The maltose/maltodextrin regulon of *Streptococcus*
727   *pneumoniae*. Differential promoter regulation by the transcriptional repressor MalR. J Biol
728   Chem. 1997;272: 30860–30865.

729   46.   DebRoy S, Gebbie M, Ramesh A, Goodson JR, Cruz MR, van Hoof A, et al.
730   Riboswitches. A riboswitch-containing sRNA controls gene expression by sequestration
731   of a response regulator. Science. 2014;345: 937–940. doi:10.1126/science.1255091

732   47.   Mellin JR, Koutero M, Dar D, Nahori MA, Sorek R, Cossart P. Sequestration of a two-
733   component response regulator by a riboswitch-regulated noncoding RNA. Science.
734   2014;345: 940–943. doi:10.1126/science.1255083

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

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

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

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

745   52.   Fu Y, Deiorio-Haggar K, Anthony J, Meyer MM. Most RNAs regulating ribosomal
746   protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria.
747   Nucleic Acids Res. 2013;41: 3491–3503. doi:10.1093/nar/gkt055

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

751 54. Turner RJ, Lu Y, Switzer RL. Regulation of the *Bacillus subtilis* pyrimidine biosynthetic
752   (*pyr*) gene cluster by an autogenous transcriptional attenuation mechanism. J Bacteriol.
753   1994;176: 3708–3722.

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

757 56. Lu Y, Turner RJ, Switzer RL. Function of RNA secondary structures in transcriptional
758   attenuation of the *Bacillus subtilis pyr* operon. Proc Natl Acad Sci USA. 1996;93: 14462–
759   14467.

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

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

764 59. Ghode P, Ramachandran S, Bifani P, Sivaraman J. Structure and mapping of spontaneous
765   mutational sites of PyrR from *Mycobacterium tuberculosis*. Biochem Biophys Res
766   Commun. 2016;471: 409–415. doi:10.1016/j.bbrc.2016.02.071

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

770 61. Slager J, Aprianto R, Veening J-W. Deep genome annotation of the opportunistic human
771   pathogen Streptococcus pneumoniae D39. Nucleic Acids Res. 2018.
772   doi:10.1093/nar/gky725

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

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

777 64. Shell SS, Chase MR, Ioerger TR, Fortune SM. RNA sequencing for transcript 5'-end
778   mapping in mycobacteria. Methods Mol Biol. 2015;1285: 31–45. doi:10.1007/978-1-
779   4939-2450-9_3

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

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

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

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

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

793   70.   Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
794   features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

795   71.   Novichkov PS, Kazakov AE, Ravcheev DA, Leyn SA, Kovaleva GY, Sutormin RA, et al.
796   RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in
797   bacteria. BMC Genomics. BioMed Central; 2013;14: 745. doi:10.1186/1471-2164-14-745

798   72.   Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE:
799   tools for motif discovery and searching. Nucleic Acids Res. 2009;37: W202–W208.
800   doi:10.1093/nar/gkp335

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

803   74.   Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate
804   alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
805   Genome Biol. BioMed Central; 2013;14: R36. doi:10.1186/gb-2013-14-4-r36

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

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

812   77.   Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
813   improvements in performance and usability. Molecular Biology and Evolution. 2013;30:
814   772–780. doi:10.1093/molbev/mst010

815   78.   Bricker AL, Camilli A. Transformation of a type 4 encapsulated strain of *Streptococcus*
816   *pneumoniae*. FEMS Microbiology Letters. 1999;172: 131–135.

817   79.   Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an
818   information aesthetic for comparative genomics. Genome Research. Cold Spring Harbor
819   Lab; 2009;19: 1639–1645. doi:10.1101/gr.092759.109

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835   **Figure legends**
836

837   **Figure 1. Schematic representation of the sequencing and data analysis methodology. (A)** A

838   description of the experimental pipeline. RNA-Seq (left column) and term-seq (middle column)

839   libraries were prepared according to protocols described in [34,35]. 5' end-Seq libraries (right

840   column) were generated by dividing the total RNA into 5' polyphosphate treated (processed) and

841   untreated (unprocessed) samples and subsequently processed according to protocols described in

842   [22,64]. White and grey lines correspond to RNA and cDNA, respectively and colored blocks

843   represent unique sequence barcodes. Illumina sequencing adaptors with index barcodes are also

844   indicated. **(B)** A brief description of the analysis pipeline. Raw reads from all three sequencing

845   methodologies obtained from three environmental conditions (early-log and mid-log growth

846   phase in the presence and absence of vancomycin) were de-multiplexed and aligned to T4

847   (NC_003028.3). Based on the reads mapped, single nucleotide coverage was calculated.

848   Coverage of the 5' end of the reads calculated for the 5' end-Seq was used to determine the

849   transcription start sites. Coverage of the 5' end of the reads from term-seq was used to determine

850   the transcription termination sites. RNA-Seq coverage was used to calculate the read-through

851   across candidate 5' untranslated regions with early transcription terminators. Read-through

852   responses of the candidates in different media conditions (rich and poor media, media with and

853   without uracil) were analyzed to identify environment-responsive RNA regulators.

854

855   **Figure 2. Genome-wide map of the *Streptococcus pneumoniae* TIGR4 transcriptional**

856   **landscape.** A map of all the transcriptional features identified. **1.** Paired-end RNA-Seq coverage

857   of the plus strand. **2.** Paired-end RNA-Seq coverage of the minus strand. **3.** Annotated operon

858   structures on the plus strand. Tile colors represent classification of the operons according to their

859   number of high confidence TSS and TTSs: 1) traditional operons consisting of multiple genes

860   with a single TSS and a TTS (green); 2) multiTSS operons consisting of multiple genes with

861   internal TSSs but one TTS (blue); 3) multiTTS operons consisting of multiple genes with a

862   single TSS but multiple internal TTSs (red); and 4) complex operons consisting of multiple

863   genes with multiple internal TSSs and TTSs (orange). To avoid clutter, simple operons

864   consisting of a single gene with a single TTS and TSS are omitted. **4.** Annotated operon

865   structures on the minus strand. **5.** 141 novel putative intergenic regulatory elements in blue. **6.**

866   742 high confidence transcriptional start sites (TSS) in green. **7.** 1864 enriched transcriptional

867   termination sites (TTS) in red. **8.** Log transformed processed/unprocessed ratio of each high

868   confidence TSS. **9.** Log transformed coverage of each predicted TTS. **10.** Positions of the known

869   structured non-coding RNAs and small RNAs. Figure was generated using Circos [79].

870

871   **Figure 3. Distribution of operon types in the genome and frequency of transcriptional**

872   **features within non-traditional operons identified using StringTie. (A)** The pie chart

873   describes the distribution of the types of operons present in T4 annotated using only high

874   confidence TSSs. A total of 388 multigene and 343 single gene operons were identified, which

875   can be divided up in 47% simple operons (single gene transcriptional units with a single TSS and

876   TTS; gray), 26% complex operons (multi-gene operons with multiple TSSs and TTSs; orange),

877   10% traditional operons (multi-gene operon with a single TSS and TTS; green), 1% multiTSS

878   operons (blue), and 16% multiTTS operons (red). (**B**) The clustered histogram describes the

879   distribution of genes and transcriptional features in non-traditional operons, where gray

880   represents the numbers of genes in the multigene operons, green represents the number of TSSs

881   within operons, red represents the number of TTSs within operons. Two-gene operons are found

882   most frequently in the non-traditional operons with one internal TSS and TTS.

883

884   **Figure 4. Variability in expression levels between genes in the same operon when grown in**

885   **rich (SDMM) and poor (MCDM) media conditions.** RNA-Seq coverage maps of complex

886   operon/regulon including the TSSs in green and TTSs in red. Size of the transcriptional features

887   represents log transformation of the processed/unprocessed ratio for TSSs and coverage for

888   TTSs. Arrows on the TSS represent the strand and direction of transcription. There are little to no

889   reads mapped to the strand opposite to coding regions. Coverage was normalized to the number

890   of uniquely mapped reads in each library and $\log_{10}$ transformed for representation (**A**) A complex

891   9-gene operon (SP_1018–SP_1026) encoding thymidine kinase, GNAT family N-

892   acetyltransferase, peptide chain release factor 1, peptide chain release factor N(5)-glutamine

893   methyltransferase, threonylcarbomyl-AMP synthase, N-acetyltransferase, serine

894   hydroxymethyltransferase, nucleoid-associated protein, and Pneumococcal vaccine antigen A

895   respectively. Genes SP_1022-SP_1026 are expressed to greater levels in MCDM than in SDMM

896   unlike genes SP_1018-SP_1021. (**B**) The maltose regulon, an example of two operons of

897   different complexities working together in response to maltose in the medium. Complex operon

898   SP_2108-SP_2112 shows greater expression in MCDM than SDMM in comparison to the

899   multiTTS operon SP_2106-SP_2017

900

901   **Figure 5. Validation of the regulatory activities of FMN, TPP riboswitches and a putative**

902   **5'-UTR regulatory candidate in different nutrient conditions.** The relative expression and

903   average RNA-Seq coverage of SP_0178 (FMN) (**A**) and SP_0716 (TPP) (**C**) increases in poor

904   (MCDM) medium compared to rich (SDMM) medium, potentially compensating for the

905   depletion of the specific ligand. Error bars represent standard error of the mean. Expression of

906   SP_0178 (FMN) (**B**) and SP_0716 (TPP) (**D**) is reduced when the poor medium is supplemented

35

907   with respective ligands thus confirming the regulatory activities of FMN and TPP riboswitches.

908   (FMN- Riboflavin; TPP- Thiamine). Expression is relative to the qPCR internal control and thus

909   for instance an expression level of 0.5 means that the expression of tested gene is 50% of the

910   expression of internal control. Error bars represent standard error of the mean across three

911   technical replicates. SP_1782 encoding ribosomal protein L11 methyltransferase **(E)** decreases in

912   poor media (MCDM) compared to rich media (SDMM).

913

914   **Figure 6. Mechanism, structure and regulatory activity of pyrR regulatory RNA element**

915   **and its mutants in the presence and absence of uracil.** (**A**) Schematic representation of the

916   proposed mechanism of regulation of *pyr* operon by pyrR RNA element. In the presence of

917   UMP, PyrR binds to the pyrR RNA and results in the formation of a premature terminator,

918   disrupting the anti-terminator formed when UMP is low, resulting in transcription termination.

919   **(B)** RNA-Seq coverage map across the *pyr* operon (SP_1278-1276) showing premature

920   transcription termination and consequently decreased expression of its genes downstream of the

921   pyrR regulator when grown in defined medium (CDM) in the presence of uracil (yellow)

922   compared to the absence of uracil (blue). Coverage was normalized to the number of uniquely

923   mapped reads in each library . TSSs are in green and the size represents the log transformation of

924   the processed/unprocessed ratio and TTSs are in red and size represents the log transformation of

925   the coverage. **(C)** qRT-PCR determining the expression of the first genes in the *pyr* operon

926   (SP_1278) in the presence and absence of uracil validates the RNA-Seq observation. Expression

927   is relative to the qPCR internal control. (**D**) Secondary structure of the *S. pneumoniae* pyrR RNA

928   regulatory element in 'off' conformation. Boxed in red and yellow are bases that were deleted in

929   M1 and M3 mutations, respectively. Bases boxed in green were replaced by indicated bases to

930   make mutation M2. Highlighted in grey are nucleotides that would base pair to form the anti-

931   terminator when the riboswitch is in the "on" conformation. (**E**) A representative qRT-PCR

932   quantification of the expression of SP_1278 transcript from pyrR RNA mutant strains cultured in

933   defined medium with or without uracil, corresponding to the regulatory activity of pyrR RNA

934   mutants. While the WTc (wild type with chloramphenicol resistance cassette) decreases

935   expression in the presence of uracil, M1 is insensitive to the ligand. M2 and M3 result in either

936   constitutive or reduced expression of the *pyr* operon. ~2 fold higher expression in M1 compared

937   to WTc in the absence of uracil could be the result of endogenous uracil having a slight

36

938   inhibiting effect on the wild type. Expression is relative to the qPCR internal control. Error bars

939   represent standard error of the mean across three technical replicates.

940

941   **Figure 7. Regulation by pyrR regulatory element is important, but not essential for *in vitro***

942   **growth of *S. pneumoniae*. (A-B)** *In vitro* growth curves of mutants when cultured in defined

943   media with (20 µg/ml) or without uracil. While mutant M2 (green) does not display a growth

944   defect, mutants M1 (orange) and M3 (maroon) have growth defects that are restored in the

945   presence of uracil indicating that a functional pyrR RNA element is important, but not absolutely

946   essential for *in vitro* growth of *S. pneumoniae*. WT (blue) does not have a growth defect in the

947   tested conditions. **(C-D)** Representative *in vitro* growth curves of mutants cultured in media with

948   (15 µg/ml) or without 5-FOA, a toxic uracil analog. All strains show varying degrees of defects

949   in the tested conditions indicating that a drug targeted against the secondary structure can

950   severely and specifically hamper growth. Mutant M3 was not included in this assay as it cannot

951   be grown without uracil. Error bars represent standard error of the mean across three biological

952   replicates.

953

954   **Figure 8. Regulation by the pyrR RNA element is crucial for *in vivo* survival and virulence**

955   **of *S. pneumoniae*.** 1x1 competition assays (mutant vs wild type T4) reveals fitness defects of

956   pyrR RNA mutants in a mouse infection model. While M1 (A) and M3 (C) display severe

957   defects in all the tested *in vivo* environments namely lung, blood and nasopharynx, M2 (B) has

958   less of a defect. Significant change in fitness (*p<0.0125)* are indicated by asterisks (*). Each data

959   point represents a single mouse.

960

961   **Supplemental figure 1. Confidence in the predicted TSSs. A.** Frequency distribution of the

962   distance between the TSS position predicted for each gene between conditions. This also takes

963   into account whether TSSs were predicted. Majority of the predictions match exactly or are very

964   close to one another between the conditions providing confidence in the TSS predictions from

965   the pooled data. **B.** Predicted TSSs show enrichment of the TATA box at the -10 and the

966   TTGACA at the -35 positions respectively as identified by MEME [72].

967

37

968   **Supplemental figure 2. Differences in the operon predictions between StringTie and**
969   **Rockhopper. (A)** StringTie reconstructed transcript grouping genes SP_0001 – SP_0014. Log
970   transformed paired-end RNA-Seq coverage map across this region. Red and Green lines
971   represent the predicted TSSs and TTSs **(B)** The pie chart describes the distribution of the types of
972   operons present in T4 identified using single-end RNA-Seq analyzed with Rockhopper. A total
973   of 474 multigene and 773 single gene operons were identified, which can be divided up in 63%
974   simple operons (single gene transcriptional units with a single TSS and TTS; gray), 29%
975   complex operons (multi-gene operons with multiple TSSs and TTSs; orange), 4% traditional
976   operons (multi-gene operon with a single TSS and TTS; green), 1% multiTSS operons (blue),
977   and 3% multiTTS operons (red). (**C**) The clustered histogram describes the distribution of genes
978   and transcriptional features in non-traditional operons described by Rockhopper, where gray
979   represents the numbers of genes in the multigene operons, green represents the number of TSSs
980   within operons, red represents the number of TTSs within operons. Two-gene operons are found
981   most frequently in the non-traditional operons with one internal TSS and TTS.

982

983   **Supplemental figure 3. Distribution and conservation of the 141 putative regulatory**
984   **candidates across 380 *S. pneumoniae* genomes. A.** Frequency distribution of the candidates
985   across the surveyed genomes. **B.** Conservation of the candidates as a measure of the mean p-
986   Distance within each candidate cluster.

987

988   **Supplemental figure 4. β-galactosidase activities of pyrR mutant strains in the presence or**
989   **absence of uracil.** The activity of the wild type pyrR decreases in the presence of uracil, while
990   M1 is insensitive to the ligand. Activity of mutant M2 results in increased Miller units even in
991   the presence of uracil. Empty *lac*Z represents a β-galactosidase reporter construct without a
992   regulatory region upstream, which was used as a negative control. Error bars represent standard
993   error of the mean across three technical replicates.

994

995

996   **Supplemental table 1. List of all the transcriptional start sites (TSSs) identified from the 5'**
997   **end-Seq in *S. pneumoniae* strain TIGR4.** The 742 high confidence primary TSSs (sheet 1), the
998   48 putative secondary high confidence TSSs (sheet 2), 1286 low confidence primary TSSs

999 identified (sheet 3), and the 74 low confidence secondary TSSs (sheet 4), their
1000 Processed/unprocessed ratios and coverages, and the lengths of the 5'-UTR thus formed are
1001 listed. Additionally, a description of whether the TSS is intergenic or within an upstream coding
1002 sequence is provided.

1003

1004 **Supplemental table 2. List of all the transcription termination sites (TTSs) identified from**
1005 **term-seq in *S. pneumoniae* strain TIGR4.** The positions of all 1864 enriched sense TTSs, and
1006 the coverage of each site pooled from all the samples are listed. Lengths of the 3'-UTRs are also
1007 listed for the sense TTSs. The predicted stem-loop structure and the number of uridines upstream
1008 of each TTS is also shown.

1009

1010 **Supplemental table 3. All the multi-gene operons identified by StringTie.** A description of
1011 the 388 multi-gene operons identified by StringTie. Number and the locus tags of genes that are
1012 in operons.

1013

1014 **Supplemental table 4. Table of candidate 5'-UTR regulatory elements.** Detailed description
1015 of the candidate 5'-UTR regulatory elements for 141 genes. Position of the TSS and its
1016 Processed/unprocessed ratio, early TTS and average coverage for each are described.
1017 Additionally, length of the 5'-UTR as well as whether it is intergenic or within the upstream
1018 coding sequence is provided.

1019

1020 **Supplemental Table 5. RNA adapters with barcode for 5' end-Seq.** List of adapters used for
1021 5' end-Seq. Highlighted in red are the barcode sequence (5'-3').
1022
1023 **Supplemental Table 6. Primers used in the study.**
1024
1025

Schematic Representation of the Protocol

A

B

A

log(Read Depth)

- MCDM (blue)
- SDMM (orange)
- Paired-end (gray)

TIGR4   963   964   965   966   967   968   Kb

SP_1018 SP_1019 SP_1020 SP_1021 SP_1022 SP_1023 SP_1024 SP_1025 SP_1026

B

log(Read Depth)

- MCDM (blue)
- SDMM (orange)
- Paired-end (gray)

TIGR4   2018   2020   2022   2024   Kb

SP_2106 SP_2107 SP_2108 SP_2109 SP_2110 SP_2111 SP_2112 F49
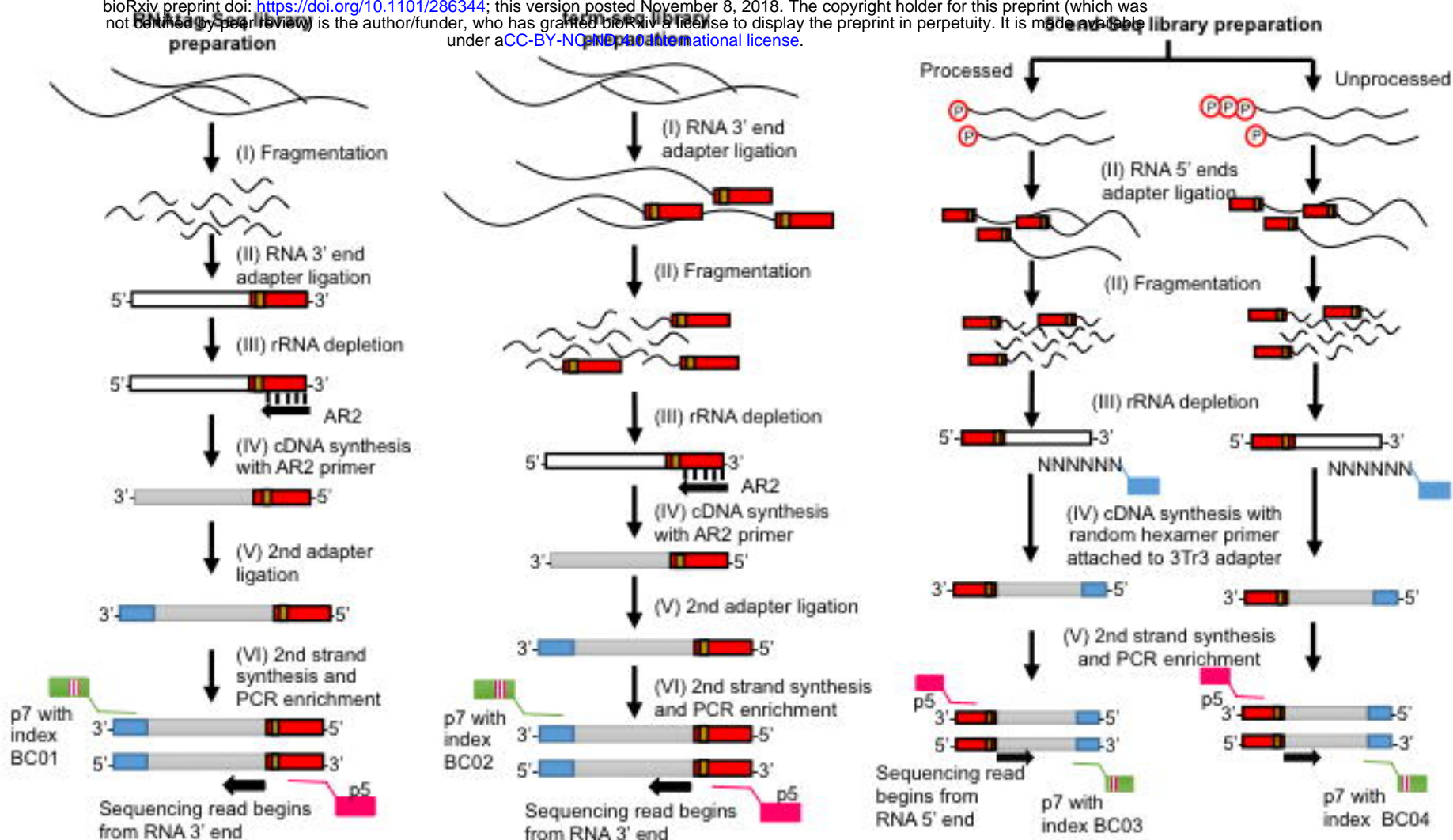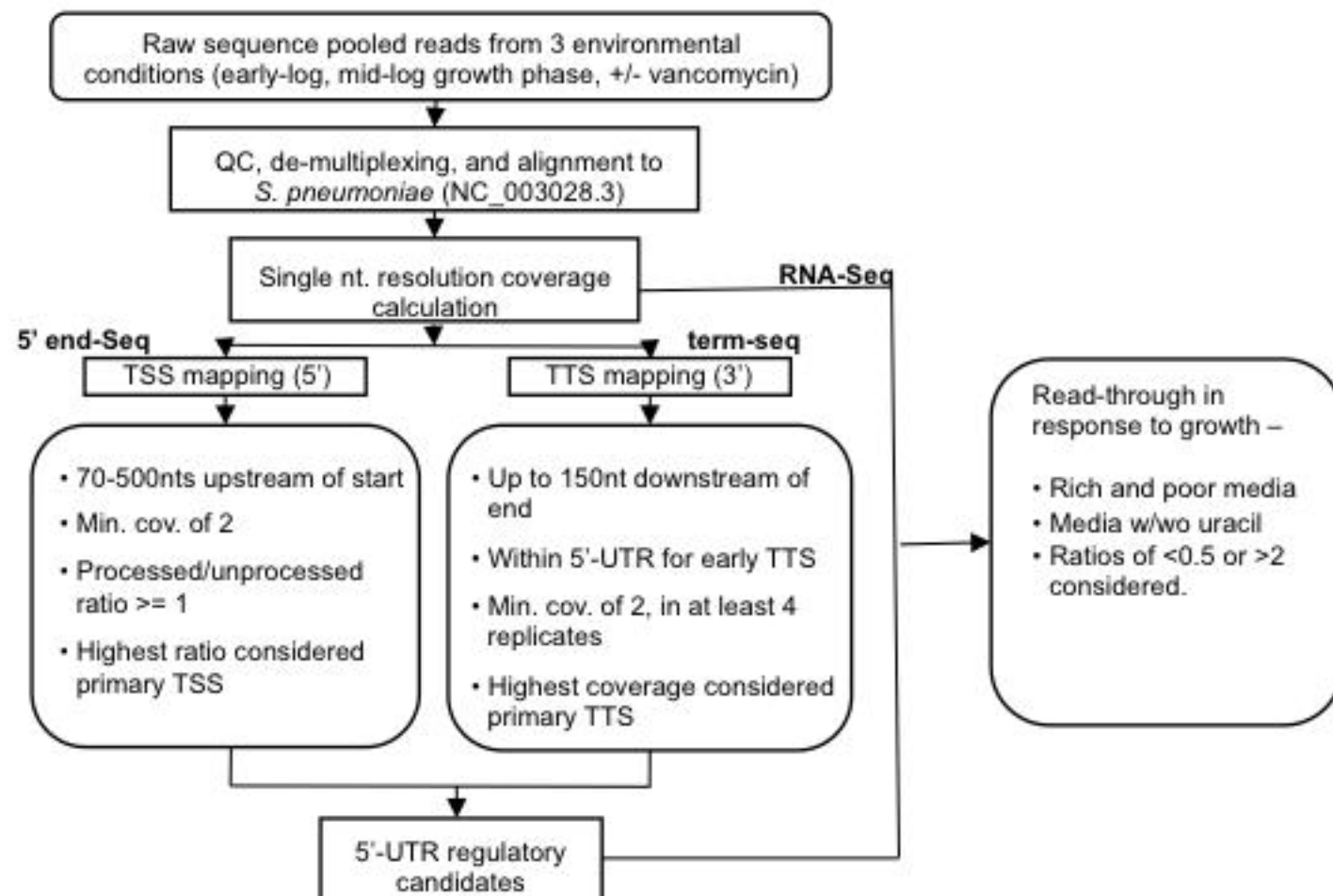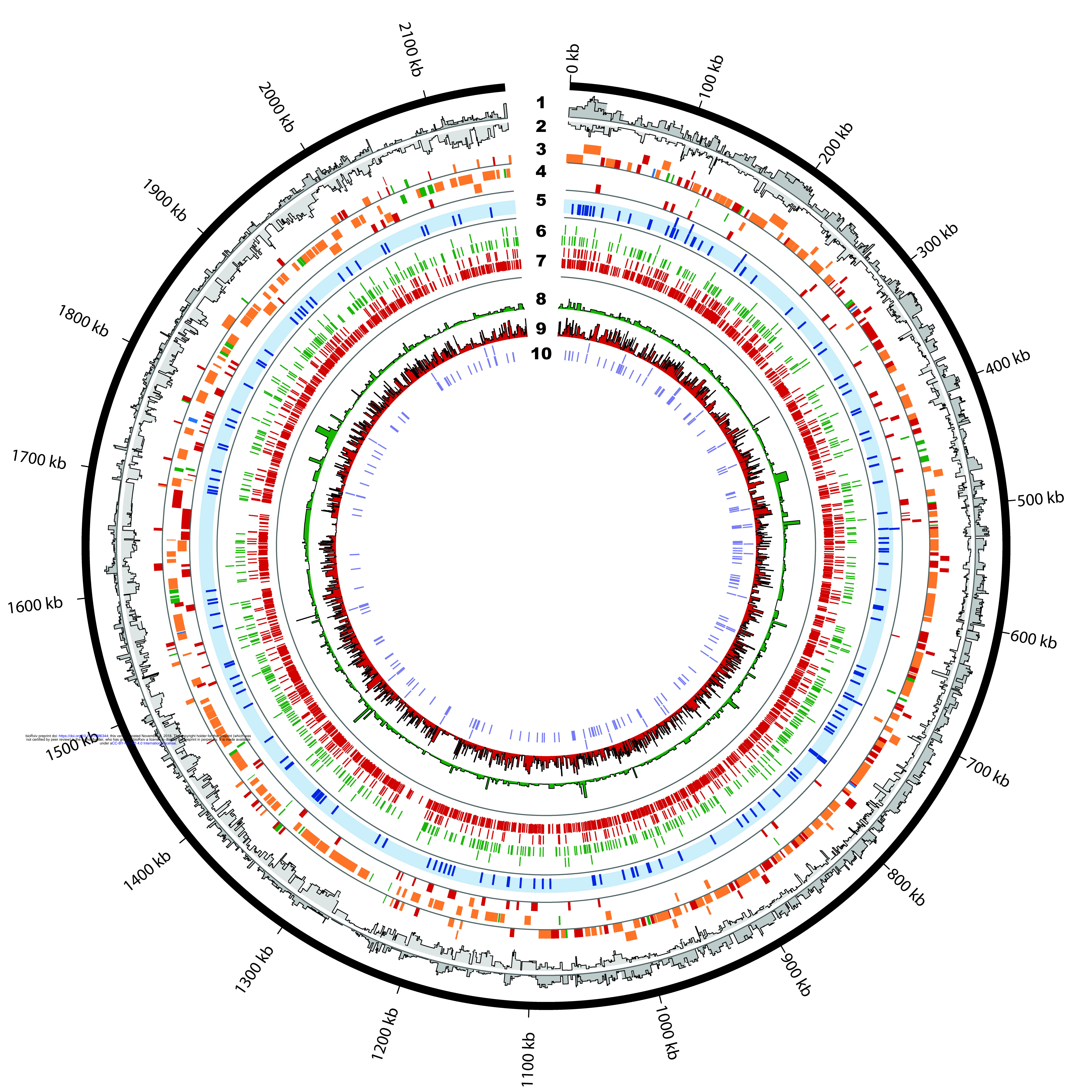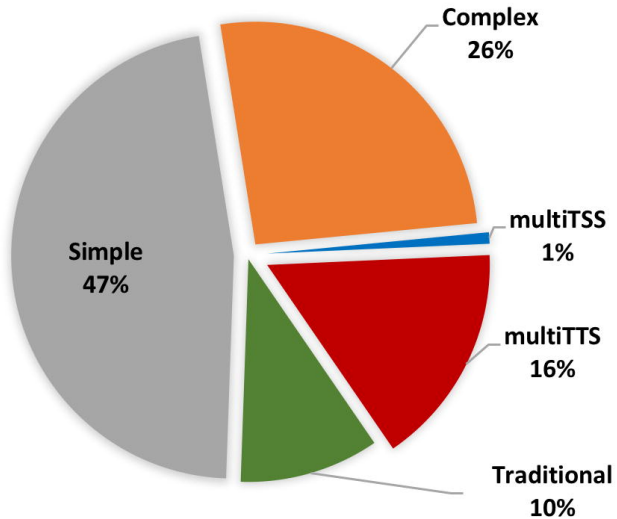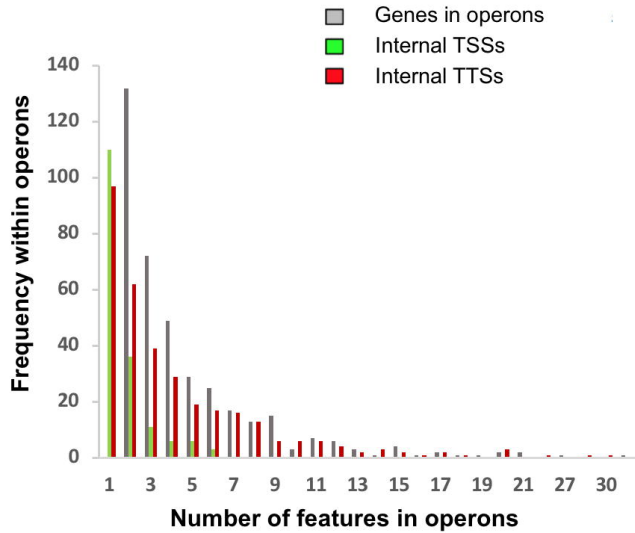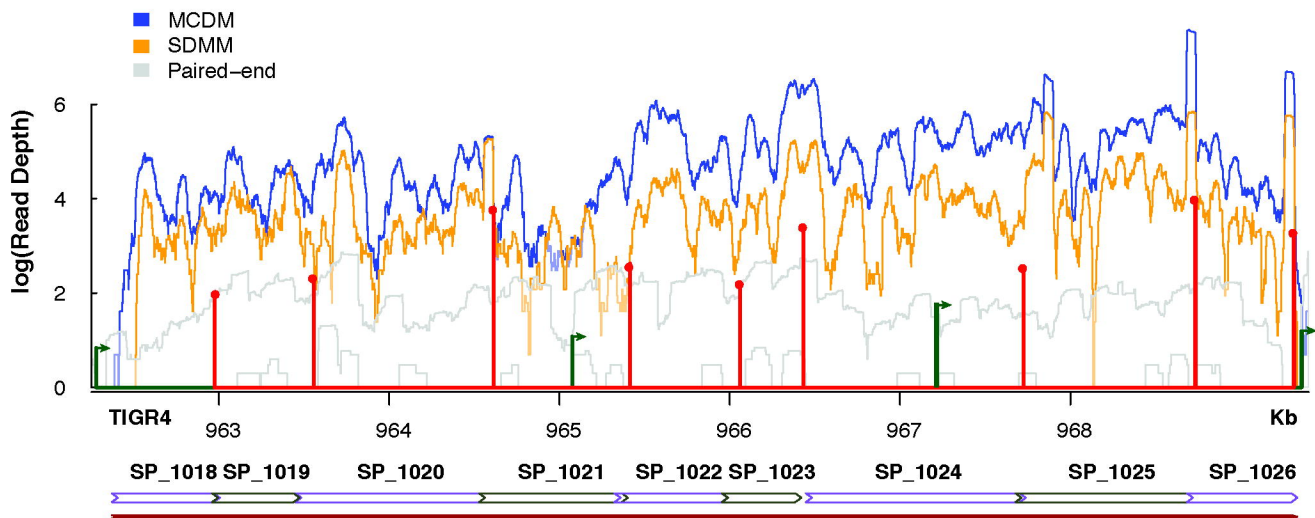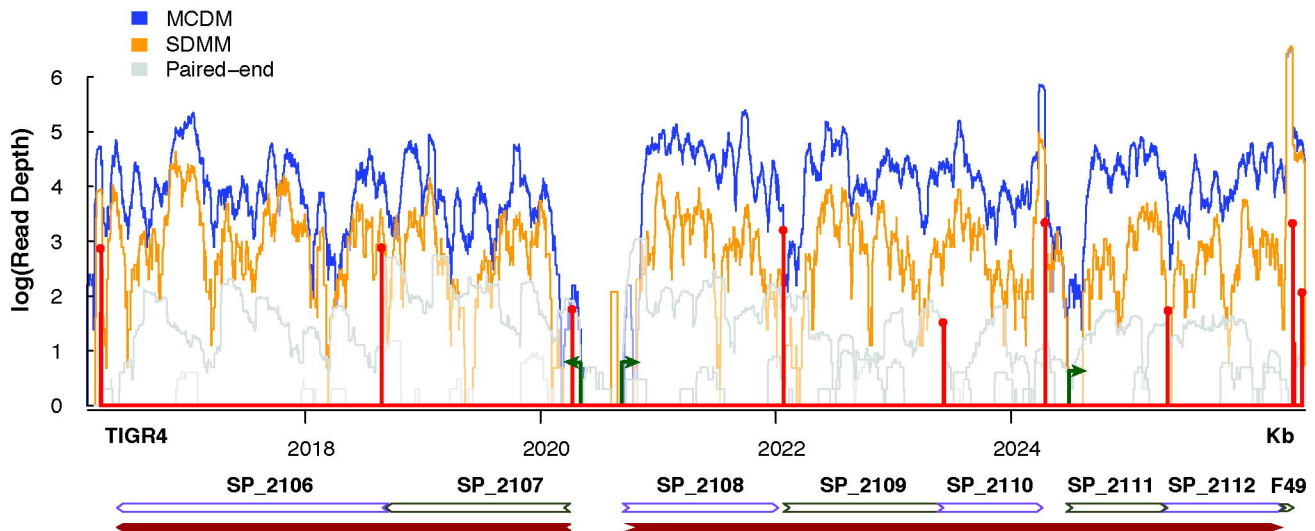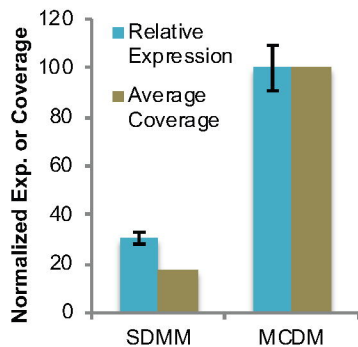
A FMN riboswitch

B FMN riboswitch
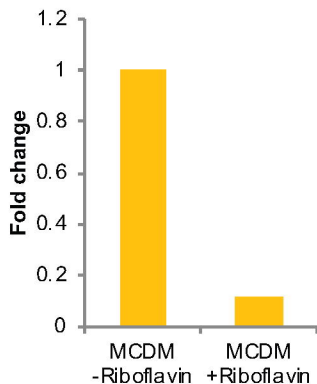
C TPP riboswitch
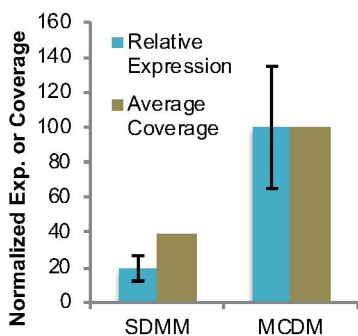
D TPP riboswitch

E SP1782

A Antitermination / Termination schematic

Low [UMP] / High [PRPP] → Antitermination

High [UMP] / Low [PRPP] → Termination

B
Read Depth
CDM -Uracil
CDM +Uracil

TIGR4   1212   1212.5   1213   1213.5   1214   Kb

SP_1276   SP_1277   SP_1278

C
Expression

CDM-Uracil   CDM+Uracil

D
M1 bases deleted
M3 bases deleted
M2
M3 bases deleted

5'-UAU␣CGCGAUAAAG␣UAAAUUUCA␣UUUCUUUU-3'

E
Expression

■ 0 µg/ml Uracil
□ 40 µg/ml Uracil

WTc   M1   M2   M3

A M1

B M2

C M3