Polycysteine-encoding leaderless short ORFs function as cysteine-responsive attenuators of operonic gene expression in mycobacteria

Running title: Attenuating sORFs coordinate a multilocus regulon

Jill G. Canestrari[1], Erica Lasek-Nesselquist[1], Ashutosh Upadhyay[1], Martina Rofaeil[2], Matthew M. Champion[2], Joseph T. Wade[1], Keith M. Derbyshire[1], Todd A. Gray[1]

[1] Division of Genetics, Wadsworth Center, New York State Department of Health, Albany, NY 12208

[2] Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN 46556

Key words:
Operon
Regulon
Cysteine
Mycobacteria
Polycysteine
Short ORFs

ABSTRACT

Genome-wide transcriptomic analyses have revealed abundant expressed short open reading frames (ORFs) in bacteria. Whether these short ORFs, or the small proteins they encode, are functional remains an open question. One quarter of mycobacterial mRNAs are leaderless, beginning with a 5'-AUG or GUG initiation codon. Leaderless mRNAs often encode unannotated short ORFs as the first gene of a polycistronic transcript. Here we show that polycysteine-encoding leaderless short ORFs function as cysteine-responsive attenuators of operonic gene expression. Detailed mutational analysis shows that one polycysteine short ORF controls expression of the downstream genes. Our data indicate that ribosomes stalled in the polycysteine tract block mRNA structures that otherwise sequester the ribosome-binding site of the 3'gene. We assessed endogenous proteomic responses to cysteine limitation in *Mycobacterium smegmatis* using mass spectrometry. Six cysteine metabolic loci having unannotated polycysteine-encoding leaderless short ORF architectures responded to cysteine limitation, revealing widespread cysteine-responsive attenuation in mycobacteria. Individual leaderless short ORFs confer independent operon-level control, while their shared dependence on cysteine ensures a collective response mediated by ribosome pausing. We propose the term ribulon to classify ribosome-directed regulons. Regulon-level coordination by ribosomes on sensory short ORFs illustrates one utility of the many unannotated short ORFs expressed in bacterial genomes.

46 INTRODUCTION

47

48 Short open reading frames (sORFs) are extremely difficult to computationally identify in
49 genome sequences; their shortened gene length approaches statistical random ORF
50 background frequencies and their amino acid sequences have limited bioinformatic value
51 (Frith et al., 2006, Crappe et al., 2013, Hemm et al., 2008, Hobbs et al., 2011).
52 Conventional mass spectrometry proteomic studies also systematically underrepresent the
53 small proteins (sproteins) encoded by sORFs, as they can be lost during sample
54 preparation or provide too few detectable peptides (Hemm et al., 2010). These limitations
55 have contributed to both a lag in (i) recognition of sORFs, and (ii) assessment of their
56 functional potential. This knowledge gap has been underscored by recent descriptions of
57 many potential novel sORFs in *Escherichia coli* and in bacteria found in the human
58 microbiome (Weaver et al., 2019, Meydan et al., 2019, Sberro et al., 2019). Previous
59 work applying complementary transcriptomic approaches to mycobacteria, in both slow-
60 growing *Mycobacterium tuberculosis*, and fast-growing *Mycobacterium smegmatis* and
61 *Mycobacterium abscessus*, suggested that their genomes contained hundreds of sORFs
62 actively producing sproteins (Miranda-CasoLuengo et al., 2016, Shell et al., 2015).
63 Ribosome profiling (Ribo-seq), together with RNA-seq and transcription start site (TSS)
64 mapping, provided genome-wide empirical evidence for the location and ribosome
65 occupancy of mycobacterial mRNAs that express sproteins (Cortes et al., 2013, Shell et
66 al., 2015, Miranda-CasoLuengo et al., 2016).

67

68 Leaderless translation is a phenomenon in which mRNAs without a 5' UTR are
69 efficiently translated, such that the first nucleotide of the start codon is the 5' end of the
70 RNA. We and others have shown that leaderless translation is abundant in mycobacteria
71 (Cortes et al., 2013, Shell et al., 2015). Many mycobacterial ORFs that are predicted to be
72 initiated by leaderless mRNAs (LL-mRNAs) are short (defined here as less than 150 nt)
73 and unannotated. The ORFs initiated at the combined transcription/translation initiation
74 start sites of LL-mRNAs are readily identifiable from transcriptomic data sets and thus
75 provide a high-confidence list of expressed novel mycobacterial sORFs. Insights into the
76 mechanistic and functional attributes of LL-mRNAs have lagged, as they are rare or
77 poorly expressed in *E. coli* but are abundant in other bacteria, archaea, and mitochondria
78 (Beck and Moll, 2018). LL-sORFs represent the first (5'-most) ORF of a transcript,
79 which positions them for a potential role in *cis*-regulation of the downstream operonic
80 genes. There is ample precedent in eukaryotes for regulation of downstream genes by
81 such "upstream ORFs" (uORFs) (Couso and Patraquim, 2017, Hinnebusch et al., 2016).
82 In prokaryotes, mechanisms have also been previously described in which uORFs have
83 been shown to regulate expression of downstream genes through a process known as
84 attenuation (Bechhofer, 1990, Henkin and Yanofsky, 2002, Oppenheim and Yanofsky,
85 1980).
86
87 Attenuation is a *cis*-regulatory mechanism often mediated by bacterial short uORFs
88 enriched in codons for the amino acid product of that biosynthetic operon. Attenuation
89 occurs when abundant charged tRNA levels allow translating ribosomes to quickly pass
90 through the modulating uORF, promoting the formation of an intrinsic terminator that

91 aborts transcription of the remainder of the operon. Low levels of charged tRNA cause
92 ribosome stalling in the uORF at codons for the end-product amino acid, facilitating the
93 formation of a competing anti-terminator structure, thereby releasing attenuation to allow
94 transcription to extend into the biosynthetic operon (Turnbough, 2019)(Fig 1). To our
95 knowledge, uORF-mediated attenuation mechanisms for cysteine have not been
96 described. A subset of predicted mycobacterial LL-sORFs conspicuously encode
97 consecutive cysteine residues, and these were found upstream of genes annotated to be
98 involved in cysteine biosynthesis (Shell et al., 2015). We hypothesized that these LL-
99 sORFs function as cysteine-sensitive attenuators.
100
101 RESULTS
102
103 We previously showed that an AUG or GUG (collectively, RUG) at the start of an
104 mRNA will direct translation of an ORF determined by that initiation codon (Shell et al.,
105 2015). Therefore, transcription start sites (TSS) beginning with RUG in mycobacteria
106 both readily identify leaderless mRNAs, and also establish the resulting translated
107 reading frames. Using published transcription start site mapping and ribosome profiling
108 fata, we identified 304 putative LL-sORFs in *Mycobacterium smegmatis* (Table S1)
109 (Martini et al., 2019, Shell et al., 2015). Cysteines that are encoded by consecutive
110 codons (polycysteine) are particularly enriched in LL-sORFs when compared to
111 annotated ORFs (Fig 2, Table S2). This enrichment of polycysteine in LL-sORFs is
112 reminiscent of consecutive tryptophans or histidines in leader sORFs in classic attenuated
113 operon models (Yanofsky, 1981).

114

115 ***Ms5788* is regulated in response to cysteine abundance**

116 One predicted LL-sORF encodes eight consecutive cysteines at its 3' end, standing out as
117 a strong candidate for a role in cysteine attenuation (Table S2). We designate this
118 polycysteine LL-sORF, *Ms5788A*, as it is immediately upstream of the annotated gene,
119 *Ms5788*, which is followed by operonic genes, including a putative thiosulfate
120 sulfurtransferase, *cysA2* (Fig 3)(Shell et al., 2015). RNA-seq and Ribo-seq profiles
121 indicate that this unannotated LL-sORF is abundantly transcribed and translated in *M.*
122 *smegmatis* (Fig S1A). Similar transcriptomic approaches in *M. tuberculosis* strongly
123 suggest that a homologous LL-sORF, designated *Rv0815A*, is expressed in that species
124 (Smith et al., 2019)(Fig S1B).

125

126 The 90 nt *Ms5788A* LL-sORF and its 97 nt 3' untranslated region (3' UTR) comprise a 5'
127 leader of gene *Ms5788* (Fig 3). This RNA leader was analyzed by RNAfold to identify
128 likely secondary structures. Nearly all of this compound leader sequence is predicted to
129 form a stable duplex structure (Fig S2). In contrast to model attenuation systems that rely
130 on competing terminator/anti-terminator stem loops of moderate stability, this predicted
131 structure appears to be exceptionally stable. Moreover, the polycysteine tract encoded at
132 the 3' end of *Ms5788A* is predicted to base-pair with the region immediately preceding
133 *Ms5788*, including the *Ms5788* Shine-Dalgarno (SD) sequence (Fig 3 and Fig S2). The
134 predicted structure of the mRNA leader led us to hypothesize that under conditions of
135 ample charged cysteine tRNA ($tRNA^{cys}$), ribosomes translating Ms5788A quickly
136 complete the process and release the mRNA at the stop codon (Fig 4A, right). As a

137 consequence, the unoccupied cysteine tract of the sORF would be available for stable
138 pairing with the SD region upstream of *Ms5788*. Under conditions of limiting cysteine,
139 however, lower levels of charged $tRNA^{cys}$ likely stall ribosomes in the polycysteine tract,
140 preventing base-pairing and stable duplex formation (Fig 4A, left). Thus, ribosome-
141 occupied sORF mRNA would not be available to base-pair with the *Ms5788* SD region,
142 leaving the latter free to recruit ribosomes for translation of *Ms5788*.
143
144 To test whether *Ms5788A* regulates *Ms5788* expression in response to changes in cysteine
145 levels, we generated a luciferase translational reporter in which a constitutive promoter
146 drives a leaderless transcript that begins at the native GUG initiation codon of *Ms5788A*
147 and continues to the AUG initiation codon of *Ms5788* (Fig 4A, top). Minimal medium
148 does not provide exogenous cysteine, requiring cells to activate the genes needed for its
149 biosynthesis. We cultured *M. smegmatis* in minimal medium, with or without cysteine
150 supplementation, and measured expression of the luciferase reporter. Luciferase activity
151 decreased for cells grown with cysteine supplementation (Fig 4Bi). Hence, we
152 hypothesized that expression of *Ms5788* is regulated in response to cysteine levels by an
153 attenuation mechanism involving the upstream LL-sORF *Ms5788A* (Fig 1B).
154
155 We next tested whether cysteine-dependent regulation of *Ms5788* requires translation of
156 the upstream *Ms5788A* LL-sORF. We mutated the GUG translation initiation codon of
157 *Ms5788A* to ACC, to prevent ribosome loading. This non-start mutation reduced
158 luciferase expression, and abolished attenuation (Fig 4Bii). These data indicate that
159 translation of *Ms5788A* is required for cysteine responsiveness, and that in the absence of

160 *Ms5788A* translation, expression of *Ms5788* is locked in an attenuated state. We also
161 tested a nonsense mutant that truncated *Ms5788A* ten amino acids before the first Cys
162 codon. This truncating mutation also reduced luciferase expression and attenuation by
163 cysteine (Fig 4Biii). The residual cysteine response of this nonsense mutant may result
164 from translation reinitiation after the stop codon. Thus, our data support a model in which
165 cysteine-responsive regulation of the downstream gene requires ribosome occupancy of
166 the LL-sORF.
167
168 We hypothesized that recoding the polycysteine tract would still allow ribosome
169 occupancy but would no longer respond to cysteine. We created an out-of-frame (OoF)
170 *Ms5788A* mutant reporter to recode the polycysteine tract to polyleucine.
171 Supplementation with cysteine had no effect on this OoF reporter, which demonstrates
172 that cysteine levels are monitored by the encoded *Ms5788A* polycysteine tract to
173 modulate *Ms5788* gene expression (Fig 4Biv).
174
175 **Attenuation requires formation of a stable RNA duplex in the *Ms5788* 5' UTR**
176 The experiments above establish that ribosome occupancy of the polycysteine tract of
177 *Ms5788A* relieves attenuated expression of the downstream reporter. Whereas the mutant
178 series in Fig 4B were predicted to have minimal effect on mRNA structure, here we
179 created mutants intended to selectively disrupt predicted duplex formation. In the first
180 series (Fig 5, ii-iv), we changed the invariant guanine in the 2nd position of cysteine
181 codons (U<u>G</u>Y) to cytosine (U<u>C</u>Y) in the last five codons of *Ms5788A* (Fig 5ii). The
182 nucleotide changes should reduce the stability of base-pair interactions with the SD

183 region, while recoding polyserine for the final five codons of *Ms5788A*. This reporter
184 exhibited constitutively elevated expression (Fig 5ii), consistent with unrestricted
185 accessibility of the SD sequence regardless of $tRNA^{cys}$ level.
186
187 To differentiate the effect of RNA duplex formation from the effect of *Ms5788A* amino
188 acid recoding, we created a mutant that only disrupted the base-pairing by changing the
189 predicted five cognate nucleotides near the SD sequence (Fig 5iii). Even though the
190 polycysteine tract remained intact, luciferase expression from this reporter was
191 insensitive to cysteine supplementation (Fig 5iii). The lack of cysteine response in this
192 mutant is consistent with the nucleotide changes disrupting base-pairing, such that
193 attenuation could not be achieved with excess cysteine. To restore base-pairing we next
194 constructed a mutant that combined the recoded *Ms5788A* and cognate non-coding
195 mutants. As expected, combining the complementary mutations reduced expression of the
196 luciferase reporter for cells grown in the presence of cysteine, consistent with restored
197 base-pairing (Fig 5iv). Interestingly, the restored response to cysteine suggests that the
198 residual four-cysteine codon content of the LL-sORF is sufficient to confer sensitivity to
199 cysteine.
200
201 In a second series (Fig 5, v-vii), we retained the full polycysteine tract and constructed a
202 silent *Ms5788A* mutant that switched the codon choice of six cysteine codons to disrupt
203 predicted base-pairing with the SD region. The elevated luciferase activity of this mutant
204 reporter was not attenuated under cysteine-replete conditions, separating the roles of
205 nucleotide and encoded amino acid sequence of *Ms5788A* (Fig 5v). We also created a

206 mutant of the predicted complementary bases near the SD. Surprisingly, this mutant
207 exhibited an attenuation response similar to wild type (Fig 5, compare i and vi). We
208 reassessed the base-pairing potential of the mRNA produced from this mutant and found
209 that new G:U base-pairings supported its folding into a stable, WT-like structure that
210 would similarly reduce SD sequence availability under conditions of ample cysteine (Fig
211 S3). Combining the *Ms5788A* and peri-SD nucleotide changes in this series resulted in an
212 expression pattern similar to that of the WT construct, consistent with the model (Fig
213 5vii).
214

215 **Cysteine-dependent regulation involving *Ms5788A* affects expression of downstream**
216 **operonic genes**

217 Classical models of ribosome-mediated attenuation invoke competing mRNA stem loops
218 that form an intrinsic terminator structure when ribosomes rapidly translate the sORF,
219 resulting in regulation at the level of transcription (Turnbough, 2019, Yanofsky, 1981)(Fig
220 1). In the case of *Ms5788* attenuation, regulation occurs at the level of translation. To
221 specifically test this, we created a reporter to assess mRNA transcription beyond the start
222 of *Ms5788*. To maintain the predicted RNA structure of the 5' leader region, an
223 independent SD sequence was added to efficiently initiate translation of transcripts that
224 extend beyond the modeled folded structure and into the luciferase gene (Fig 6). The added
225 15 nt sequence does not include identifiable complicating transcriptional or translational
226 sequences. Luciferase activity of this reporter was insensitive to cysteine, and expression
227 levels were consistently high (Fig 6ii). These data indicate that *Ms5788* attenuation does
228 not result from transcription termination. Translational repression can indirectly affect

229 transcription over longer distances due to polarity that is caused by Rho-dependent
230 transcription termination and/or by enhanced RNase processing of the untranslated RNA
231 (Deana and Belasco, 2005, Martini et al., 2019, Peters et al., 2011). To determine if
232 translational repression of *Ms5788* by attenuation leads to polar effects on downstream
233 genes, we constructed a translational fusion that extends to *Ms5789* (*cysA2*). This more
234 distal site, 516 nt 3' of the *Ms5788* start, exhibited cysteine responsiveness (Fig 6iii).
235 Therefore, the ambient cysteine response imparted by *Ms5788A* to *Ms5788* translation
236 initiation is also reflected in *Ms5789* gene expression. Taken together, our data support a
237 model in which *Ms5788* expression is regulated by translational attenuation through
238 *Ms5788A*-controlled SD availability, while *Ms5789* is likely regulated by polarity, due to
239 the absence of elongating ribosomes in *Ms5788*.
240
241 **Cysteine-dependent regulation of *Ms5789* and *Ms5790* by *Ms5788A* occurs in a**
242 **chromosomal context**
243
244 To assess whether our reporter-supported model is valid in the context of the native
245 locus, we performed label-free quantitative (LFQ) mass spectrometry-based proteomic
246 analysis to determine differences in protein expression for cells grown with or without
247 cysteine supplementation. Whole cell extracts were prepared from cultures of *M.*
248 *smegmatis* grown in minimal media +/- cysteine supplementation. Tryptic digests of
249 whole cell lysates were subjected to nanoUHPLC-MS/MS. Peptides were identified and
250 quantitated using LFQ based peak integration (Cox and Mann, 2008, Bosserman et al.,
251 2017). As expected, the abundance of most proteins is unchanged between the two

252 conditions (+/- cysteine), reflected in the linear correlation on the diagonal of the scatter
253 plot (Fig 7A). Proteins below the diagonal were more abundant in cells grown without
254 cysteine supplementation. The small (159 AA) and hydrophobic Ms5788 was not
255 detected in these experiments. However, *Ms5789* and *Ms5790* are predicted to be co-
256 transcribed in an operon with *Ms5788* (Martini et al., 2019), and expression of the
257 encoded proteins Ms5789 and Ms5790 was higher in cells grown without cysteine (Fig
258 7A, diamonds 1 and 2), consistent with release from attenuation through *Ms5788A*.
259
260 To test the hypothesis that polycysteines in *Ms5788A* control Ms5789 and Ms5790
261 expression, we generated an out-of-frame (OoF) mutation in *Ms5788A* by deletion of a
262 single nucleotide from the chromosomal locus. This deletion shifts the reading frame to
263 encode valines and alanines in place of the polycysteine tract, and it adds six additional
264 amino acids (ERSRAL) prior to encountering a stop codon (Fig 7B), while preserving the
265 potential for nearly wild-type RNA duplex formation. This mutant exhibited lower
266 Ms5789 and Ms5790 expression levels that were unaffected by cysteine supplementation
267 (Fig 7B, diamonds move into the diagonal). This result links a point mutation in
268 *Ms5788A* with gene expression of *Ms5789* and *Ms5790*, functionally demonstrating the
269 polycistronic structure of this operon that is controlled by a LL-sORF.
270
271 Our working model suggests that if *Ms5788A* directs attenuation, its deletion would
272 elevate operon expression of the downstream genes. We created a precise chromosomal
273 deletion of *Ms5788A* and subjected this mutant to LFQ proteomics. As predicted,
274 Ms5789 and Ms5790 were no longer responsive to ambient cysteine supplementation,

275 appearing in the diagonal of unresponsive genes (Fig 7C, diamonds 1 and 2 now in the
276 diagonal). Moreover, absolute levels of Ms5789 and Ms5790 were higher in the
277 Δ*Ms5788A* strain than in either the wild-type or the OoF mutant, regardless of cysteine
278 supplementation. This indicates that the *Ms5788A* deletion leaves the *Ms5788* SD
279 sequence fully available for canonical translation initiation, and results in an elevated
280 expression of the operonic Ms5789 and Ms5790 (Fig 7B, C diamonds 1 and 2). Thus,
281 data from wild-type and targeted chromosomal mutant *M. smegmatis* are in full
282 agreement with our reporter-generated data and strengthen the conclusion that the
283 *Ms5788A* LL-sORF modulates operonic gene expression through a polycysteine-relieved
284 attenuation mechanism.
285
286 **Widespread cysteine-dependent regulation in *M. smegmatis* associated with cysteine-**
287 **rich LL-sORFs**
288
289 While our in-depth analysis focused on the LL-sORF with the longest polycysteine tract,
290 we hypothesized that other polycysteine LL-sORFs also regulate operonic gene
291 expression in response to cysteine levels. We identified six more LL-sORFs that contain
292 at least two consecutive cysteine codons and are located upstream of annotated genes. We
293 then examined our LFQ proteomic data for the protein levels encoded by putative
294 operonic genes downstream of these LL-sORFs for cells grown with/without cysteine
295 supplementation. Remarkably, detectable proteins encoded by operons led by
296 polycysteine LL-sORFs were also upregulated in low cysteine, falling below the diagonal
297 (Fig 7A, black circles). Similar to *Ms5788A* controlling operonic expression of Ms5789

298 and Ms5790, these other LL-sORFs appear to control the expression of their operonic
299 genes in response to cysteine supplementation. In cysteine-replete medium, expression of
300 some of these proteins (Ms0113, Ms0934, Ms4527, Ms5279, and Ms5280) was below the
301 threshold of detection for reliable quantification [$<10^5$ LFQ intensity (a.u.)], indicative of
302 tight attenuation. In the low-cysteine environment of minimal medium, attenuation at
303 these polycysteine LL-sORF loci is apparently relieved, resulting in increased production
304 of the operon-encoded proteins. The polycysteine LL-sORFs exhibit RNA-seq and Ribo-
305 seq expression profiles consistent with their robust expression during cysteine replete
306 growth (Figs S1 and S4). None of these expressed LL-sORFs were identified by genome
307 annotation pipelines. Each of these cysteine-responsive operons contains genes annotated
308 for cysteine associated activities (Table 1). Collectively, these data reveal a cysteine-
309 metabolic regulon, whose concerted response is controlled independently at each locus by
310 an expressed LL-sORF that includes consecutive cysteine codons.
311
312 LFQ proteomics does not comprehensively identify all proteins in a proteome. We looked
313 for additional loci with the same hallmarks of responsive attenuation: an expressed
314 polycysteine LL-sORF upstream of annotated operonic genes. *Ms4536A* is a transcribed
315 polycysteine-encoding LL-sORF that was identified upstream of a single gene whose
316 protein product, Ms4536/TQXA, was not detected by proteomics (Table 2, Tables S1 and
317 S3, Fig S4E). Without corroborating indicators of cysteine response or function of the
318 encoded operon protein, we can only speculate at its membership in the *M. smegmatis*
319 cysteine LL-sORF regulon. The independent evolution of LL-sORFs makes it unlikely
320 that our experimental reference species, *M. smegmatis*, contains all of the LL-sORF

321 operons in the mycobacterial pan-genome. In one example, transcriptomic profile data for
322 *M. tuberculosis* clearly identified *Rv2334A* as an unannotated gene meeting all of the
323 criteria demonstrated in the *M. smegmatis* regulon operons: an expressed LL-sORF with
324 a C-terminal polycysteine tract, followed by genes annotated as *cysK1* and *cysE* (Table 2,
325 Fig S4G).
326
327 The availability of complete genome sequence information on diverse mycobacteria
328 provided an opportunity to track the evolution of these polycysteine sORFs. We searched
329 the genomes of 41 mycobacterial species using complementary approaches predicated on
330 the sequence of each *M. smegmatis* or *M. tuberculosis* LL-sORF. The inherently low
331 informatic content of sORFs make them poor BLAST queries, so we also used the
332 position of the flanking annotated orthologous genes as landmarks. We identified
333 genomic sequences consistent with conservation of the LL-sORFs expressed in *M.*
334 *smegmatis* and *M. tuberculosis* (Table S3). The distribution of the conserved LL-sORFs
335 is summarized as barcodes adjacent to each species on the *Mycobacterium* genus
336 phylogenetic tree (Fig 8A). Sequence logos derived from multiple alignments of the
337 amino acid sequences encoded by the LL-sORFs further support the evolutionary
338 selection of their consecutive cysteine tracts (Fig 8B). The similarity between Ms4536A
339 and Rv2334A strongly suggests homology by common origin, yet the context, including
340 operon genes, is not homologous, suggesting origin by rearrangement rather than a
341 merodiploid duplication or horizontal gene transfer event. Non-cysteine amino acids are
342 also conserved in some LL-sORFs, suggesting that they provide function through
343 ribosomal interaction, or support activities as an independent small protein product, or are

344 encoded by codons that are constrained at the nucleotide level.

345

346 DISCUSSION

347

348 **An attenuation mechanism that controls translation in response to amino acid**
349 **availability**

350 In the work presented here, we demonstrate attenuation of a cysteine biosynthesis pathway
351 locus. Attenuation is a recurring theme in biosynthetic pathways for nucleosides and amino
352 acids, in which the end product of the pathway interacts with the 5' leader of the
353 biosynthesis-encoding transcript to reduce (attenuate) expression of the operonic ORFs
354 downstream (Turnbough, 2019). However, these models typically involve the formation of
355 competing alternate hairpin structures that function as an intrinsic terminator when the end
356 product is plentiful (Fig 1A). By contrast, the *Ms5788A* LL-sORF featured here defines a
357 class of translational attenuator that indirectly assesses charged $tRNA^{cys}$ availability to
358 modulate expression of the downstream operonic genes. Blocking ribosome access to the
359 SD sequence is a translational attenuation mechanism reported for sORF control of
360 macrolide antibiotic resistance operons (Fig 1B) (Ito and Chiba, 2013). In the exit tunnel
361 of the elongating ribosome, erythromycin interacts with the nascent peptide of the ermCL
362 sORF of *Staphylococcus aureus*; the antibiotic-arrested ribosome alters mRNA hairpin
363 structures to free the SD, allowing translation initiation of the *ermC* gene (Arenz et al.,
364 2014). Chloramphenicol has a similar effect on the *cat* operon in *Bacillus subtilis* (Gu et
365 al., 1994). These shared, yet varied, mechanisms illustrate how ribosome health is actively
366 assessed and remedied, whether the cause is antibitioic attack or amino acid starvation.

367

368 **The need for cysteine regulation in mycobacteria**

369 Why might mycobacteria need a multi-locus cysteine regulon? Mycobacteria rely on a
370 cysteine derivative, mycothiol (MSH), to modulate redox balance (Loi et al., 2015, Xu et
371 al., 2011). The *mshA* gene (*Ms0933*) encodes the first enzyme in MSH biosynthesis, and it
372 resides in an operon with the hallmarks of cysteine attenuation (Fig 7A, Fig S4B). The
373 multiple roles of cysteine in mycobacterial protein synthesis, redox and sulfur metabolism
374 likely require subtle, independent fine-tuning of the enzymes involved in these respective
375 pathways. The slight differences in cysteine composition and architectures of the
376 polycysteine LL-sORFs could impart varying mechanisms as well as customized levels of
377 operon gene expression.

378

379 **Attenuating sORF requirements**

380 What are the cysteine codon requirements of an effective small ORF attenuator? Our
381 criterion that LL-sORFs encode two consecutive cysteine codons may seem to be a low
382 threshold, yet *cis*-encoded proteins only detected in unsupplemented minimal medium
383 (baseline of Fig 7A) demonstrate the effectiveness of two consecutive cysteines in LL-
384 sORFs in relieving attenuation. Each of the five LL-sORFs encoding at least three
385 consecutive Cys codons (Table S2) were found to direct a cysteine response of their
386 operonic proteins. Our *Ms5788A* LL-sORF analysis demonstrated that base-pairing was
387 needed to impose attenuation, but it is premature to speculate that similar structures are
388 required at all responsive loci. The additional cysteine-responsive operons that we identify
389 are all led by an expressed polycysteine LL-sORF, but whether each attenuation

390 mechanism is transcriptional or translational has not yet been determined. Regulation of
391 transcription initiation in response to cysteine levels does not appear to be a factor at the
392 *Ms5788* locus; the promoter remained intact in the *Ms5788A* mutants in Fig 7B, C,
393 indicating that all of the cysteine response at this locus is directed by the attenuation
394 mechanism we detailed and not via transcriptional activation. RNA-seq profiles of the LL-
395 sORFs presented here (Fig S1, S4) are also consistent with robust constitutive transcription
396 in the cysteine-replete medium used for routine culturing.
397
398 **Evolution of a cysteine attenuation regulon; the Cys ribulon**
399 The independent evolution of similar polycysteine LL-sORF architecture associated with
400 cysteine attenuation at multiple loci in mycobacteria indicates that coordinated expression
401 is important, and that LL-sORF-directed attenuation is effective. The co-regulation of
402 individual operons, caused by ribosome stalling at sORFs that are cysteine sensors in this
403 scenario, functionally defines a regulon that we term the "ribulon" (Fig 9). The term ribulon
404 reflects that control is dictated by the ribosome pausing and not a protein *per se*, and it
405 would be functionally analogous to regulons controlled by dedicated DNA-binding
406 transcription factors. We speculate that the evolution of regulation by attenuation is
407 simplified in mycobacteria by the robust nature of leaderless translation, and that the
408 evolution of a dedicated transcription factor and its cognate binding sites in the promoters
409 of target operons is more problematic than exploiting LL-sORFs in a genus that exhibits
410 frequent and robust LL-mRNA expression. As the first genes in their transcripts, LL-
411 mRNAs are ideally positioned to *cis*-regulate expression of downstream genes.
412 Additionally, transcription start sites are preferentially associated with purines (R), and the

413 +2 transcript position is preferentially associated with pyrimidines (Y) (Martini et al.,
414 2019). Thus, transcription often begins at RYN trinucleotide sequences. Our previous study
415 showed that a 5' RUG trinucleotide confers robust leaderless translation initiation (Shell et
416 al., 2015), so many transcription start sites are only one or two changes from also initiating
417 translation. It is not yet clear whether leaderless architecture *per se* offers advantages for
418 attenuation and may have been selected over canonical Shine-Dalgarno translation
419 initiation, or whether the sole criterion of an RUG sequence at the transcription start site is
420 simply a relatively modest requirement. It is clear, however, that in mycobacteria,
421 polycysteine LL-sORFs integrate two levels of coordination: locally by modulating
422 polycistronic operon gene expression, and globally by coordinating the response of
423 cysteine-related operons.
424
425 **Concluding remarks**
426 Given the prevalence of LL-sORFs in mycobacteria, we speculate that other translational
427 regulons have evolved as an alternative to transcriptional regulons controlled by DNA-
428 binding transcription factors. LL translation is considered to be the ancestral form of
429 ribosome delivery (Duval et al., 2013, Nakamoto, 2009, Zheng et al., 2011), indicating
430 that LL-sORF regulons may be ancient and widespread. We propose that LL-sORFs
431 define a functional subclass of small *de novo* genes that effectively decode translational
432 stress into broad regulatory effects.
433
434 EXERIMENTAL PROCEDURES
435

436 Bacterial strains and culture

437 *M. smegmatis* wild-type mc^{2}155 and its derivatives were grown in tryptic soy broth +
438 0.05% Tween 80 (TSBT) or on TSA plates, and cultured at 37 °C. Antibiotic selection
439 for reporter maintenance or mutation selection strategies included apramycin (12.5 μg ml^{-}
440 ^{1} on agar, 10 μg ml^{-1} in broth), hygromycin (100 and 25 μg ml^{-1}), kanamycin (50 and 10
441 μg ml^{-1}), and zeocin (50 and 25 μg ml^{-1}).

442

443 For the cysteine attenuation study, bacteria were cultured in minimal media. Base
444 medium per liter: 6 g Na_2HPO_4 (anhydrous), 3 g KH_2PO_4, 0.5 g NaCl, 1 g NH_4Cl, 0.05%
445 (v/v) Tween-80. After autoclaving, 0.2 % glucose and micronutrients were added to final
446 concentrations: $MgSO_4$ to 1 mM, $CaCl_2$ to 100 μM, H_3BO_3 to 4x10^{-7} M, $CoCl_2.6H_2O$ to
447 3x10^{-8} M, $CuSO_4.5H_2O$ to 1x10^{-8} M, $MnCl_2.4H_2O$ to 8x10^{-8} M, $ZnSO_4.7H_2O$ to 1x10^{-8}
448 M, $FeSO_4.7H_2O$ to 1x10^{-6} M. L-cysteine or L-cystine (exogenously stable dimeric
449 cysteine) was supplemented at 200 μg ml^{-1} as noted.

450

451 Luciferase assays

452 Reporters were generated by long-primer-dimer PCR to recreate the LL-sORF and leader
453 sequence of *M. smegmatis Ms5788*. PCR products were cloned by Infusion (Clontech)
454 and verified by DNA sequence analysis. The NanoLuc (Promega) luciferase ORF is
455 carried on a plasmid that confers apramycin resistance and integrates at the L5 *attB* site in
456 mycobacteria. Mycobacterial cultures were grown in minimal medium for luciferase
457 assays. Luciferase activity in a culture was assessed by the addition of NanoGlo
458 (Promega) substrate and then measuring luminescence normalized to culture density.

459

460 Chromosomal mutants of *M. smegmatis*

461 A precise deletion of *Ms5788A* was created using a targeting plasmid that integrates via a
462 single cross-over, allowing selection of a hyg^r intermediate and, then, is resolved by a
463 second homology-driven event in which a *sacB*/*galK* counter-selection allows enrichment
464 for the deleted recombinant (Barkan et al., 2011). The OoF point mutant was created by a
465 recombineering approach that used a single-stranded oligonucleotide template to
466 introduce an additional adenine in *Ms5788A* (van Kessel and Hatfull, 2008). Co-
467 electroporation of 2 x 10^{-10} mol of 60-mer oligo with 500 ng of an episomal zeo^r plasmid
468 (pGE324, Zeo-*sacB*) allowed zeocin selection and isolation of the electrocompetent
469 population of *M. smegmatis*. Mismatch-sensitized PCR (MAMA-PCR) screening (Cha et
470 al., 1992, Swaminathan et al., 2001) of isolates identified clones that integrated the
471 additional adenine. Deletion and point mutants were verified by genomic DNA PCR and
472 sequencing.

473

474 Mass spectrometry

475 Wild-type and mutant derivative *M. smegmatis* were cultured in minimal media with or
476 without cysteine supplementation, harvested by centrifugation and cryo-milled (Retsch
477 MM400, Haan, Germany) for mass spectrometry analysis. Reagents were of LC-MS
478 quality or higher and obtained from Sigma Aldrich unless indicated. Milled cell pellets
479 were digested with trypsin using commercial S-Traps (Protifi, NY). Briefly 50 μg of
480 protein from each milled cell pellet was re-suspended in 6% SDS 10 mM Tris-2-Carboxy
481 ethyl phosphine in 100 mM tri ethyl ammonium bicarbonate (TEAB), heated at 95 °C for

482 3 minutes, then alkylated with 10 mM iodoacetamide in the dark for 15 min. Samples
483 were acidified by addition of H_3PO4_4 to 1.2% final concentration (v/v) and flocculated by
484 7-fold addition of 95:5 MeOH:100 mM TEAB prior to collection on an S-trap column
485 (Zougman et al., 2014). Washing and conditioning was preformed three times by addition
486 of 150 µ MeOH buffer, as described above. 1 µg of sequencing-grade trypsin (Promega,
487 WI) was added to each sample and digested at 37 °C for 8 hours. Peptides were isolated,
488 acidified and desalted using Stage tips packed into P100 pipette tips, and dried using a
489 MiVac (Genvac UK) prior to LC-MS/MS analysis (Rappsilber et al., 2003).
490
491 NanoUHPLC-MS/MS was performed essentially as described (Bosserman et al., 2017,
492 Bosserman et al., 2019). 1 µg of each digest was analyzed in technical triplicate and
493 biological duplicate on an Orbitrap instrument running a TOP15 data-dependent
494 acquisition (Q-Exactive Thermo San Jose, CA). Protein spectral matching and Label Free
495 quantification were performed using MaxQuant (Cox and Mann, 2008) against the *M.*
496 *smegmatis* FASTA combined with contaminants from the Uniprot database, LFQ param
497 (UP000000757 6,595 entries). LFQ parameters were set to default, quantification was
498 restricted to proteins >2 peptides (missing peaks enabled). Target-decoy was used to
499 determine False Discovery Rates (Elias and Gygi, 2007) and proteins at a global 1% FDR
500 were used for quantification. Data reduction and significance testing were performed
501 using a modified LIMMA methodology (Efstathiou et al., 2017). Protein search and
502 RAW data files are accessible at the Center for Computational Mass Spectrometry via
503 ftp://MSV000084381@massive.ucsd.edu (Deutsch et al., 2017). Proteins quantitatively

504 replicated in one condition but absent in another, were given an arbitrary intensity of
505 $5x10^4$ (below detection threshold) for ease in visualization.
506
507 Phylogenetic analysis
508 The genomes and proteomes for *Mycobacterium abscessus* subsp. massiliense
509 (NC_018150.2), *M. africanum* strain 25 (CP010334,1), *M. litorale* strain F4
510 (CP019882.1), *M. avium* 104 (NC008595.1), *M. ulcerans* Agy99 (NC_008611.1), *M.*
511 *vanbaalenii* PYR-1 (NC_008726.1), *M. marinum* M (NC_010612.1), *M. liflanddii*
512 128FXT (NC_020133.1), *M. kansasii* ATCC 12478 (NC_022663.1), *M. gilvum* Spyr1
513 (NC_014814.1), *M. bovis* AF2122/97 (NC_002945.4), *M. tuberculosis* H37Rv
514 (NC_000963.3), *M. sinense* strain JDM601 (NC_015576.1), *M. canettii* CIPT 140010059
515 (NC_015848.1), *M. chubuense* NBB4 (NC_018027.1), *M. intracellulare* MOTT-64
516 (NC_016948.1), *M. intracellulare* ATCC 13950 (NC_016946.1), *M. neoaurum* VKM ac-
517 1815D (NC_023036.2), *M. haemophilum* ATCC 29548 (NZ_CP0118883.2), *M. simiae*
518 ATCC 25275 (NZ_HG315953.1), *M. goodii* strain X7B (NZ_CP012150.1), *M. fortuitum*
519 strain CT6 (NZ_CP011269.1), *M. phlei* strain CCUG 21000 (NZ_CP014475.1), *M.*
520 *immunogenum* strain CCUG 47286 (NZ_CP011530.1), *M. chelonae* CCUG 47445
521 (NZ_CP007220.1), *M. vaccae* 95051 (NZ_CP011491.1), *M. chimaera* strain AH16
522 (NZ_CP012885.2), *M. caprae* strain Allgaeu (NZ_CP016401.1), *M. colombiense* CECT
523 3035 (NZ_CP020821.1), *M. dioxanotrophicus* strain PH-06 (NZ_CP020809.1), *M.*
524 *marseillense* strain FLAC0026 (NZ_CP023147.1), *M. lepraemurium* strain Hawaii
525 (NZ_CP021238.1), *M. shigaense* strain UN-152 (NZ_AP018164.1), *M. stephanolepidis*
526 (NZ_AP018165.1), *M. pseudoshottsii* JCM 15466 (NZ_AP018410.1), *M. paragordonae*

527 49061 (NZ_CP025546.1), *M. rutilum* strain DSM 45405 (NZ_LT629971.1), *M.*
528 *thermoresistibile* strain NCTC10409 (NZ_LT906483.1), *M. hassiacum* DSM 44199
529 (NZ_LR026975.1), and *M. microti* strain 12 (CP010333.1) were downloaded from NCBI.
530 Orthologs of *M. smegmatis* proteins were identified by a reciprocal best BLAST hit
531 approach and aligned in MAFFT v.7.058b (Katoh and Standley 2013). A supermatrix
532 was created from the concatenation of 1560 of these protein alignments with an in-house
533 Python script. Alignments included in the supermatrix were required to have 40-41 of
534 the 41 mycobacterial species present. IQ-TREE v.1.6.9 (Nguyen et al., 2015) generated a
535 maximum likelihood (ML) phylogeny under an LG+R4 model of evolution (Soubrier et
536 al., 2012; Yang 1995; Le and Gascuel 2008). Support values were generated by the
537 ultrafast bootstrap method with 1000 replicates (Minh et al., 2013). The ML tree was
538 visualized in FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).
539
540 Identification of sORFs in additional mycobacterial genomes
541 The orthologs to all *M. smegmatis* ORFs with an upstream sORF were identified in other
542 mycobacteria via a reciprocal best blast hit approach and the regions 1000 nucleotides
543 upstream of these genes were extracted. *M. smegmatis* sORF proteins were queried
544 against these upstream regions via tBLASTn searches to identify orthologous sequences.
545 The coordinates for these sequences were obtained from the blast results and extended to
546 the match the length of the corresponding *M. smegmatis* sORF. To ensure that we
547 captured the start and stop codon, we extended these coordinates by an additional 5-10
548 amino acids on both 5' and 3' ends. These coordinates and the strand information
549 captured from the blast results were used to extract and translate sORFs from 1 Kb

upstream regions with an in-house Python script. All sORFs were visually inspected and initially aligned in MEGA v. 70.26 (Kumar et al., 2016). In some cases, additional sORF orthologs were identified by predicting proteins (anything between two stop codons) in upstream 1 Kb regions with the getorf subroutine from EMBOSS v. 6.3.1 (Rice et al., 2000).

Multiple alignments of deduced LL-sORF amino acid sequences were performed using a web-based tool and default clustal algorithm (Madeira et al., 2019). The aligned sequences were compiled into a sequence logo infographic (weblogo.berkeley.edu).

Acknowledgments

This work was supported by NIH grants R21AI147608, R21AI119427, R01AI097191, R21AI117158 to KMD, TAG, JTW, and MMC, and by Wadsworth Center Cores for Bioinformatics & Statistics, Applied Genomic Technologies, and Media & Tissue Culture. We thank Dr. Paul Masters for his thoughtful comments in manuscript preparation.

REFERENCES

ARENZ, S., MEYDAN, S., STAROSTA, A. L., BERNINGHAUSEN, O., BECKMANN, R., VAZQUEZ-LASLOP, N. & WILSON, D. N. 2014. Drug sensing by the ribosome induces translational arrest via active site perturbation. *Mol Cell,* 56, 446-52.

BARKAN, D., STALLINGS, C. L. & GLICKMAN, M. S. 2011. An improved counterselectable marker system for mycobacterial recombination using galK and 2-deoxy-galactose. *Gene,* 470, 31-6.

BECHHOFER, D. H. 1990. Triple post-transcriptional control. *Mol Microbiol,* 4, 1419-23.

577 BECK, H. J. & MOLL, I. 2018. Leaderless mRNAs in the Spotlight: Ancient but
578 Not Outdated! *Microbiol Spectr,* 6.
579 BOSSERMAN, R. E., NGUYEN, T. T., SANCHEZ, K. G., CHIRAKOS, A. E., FERRELL,
580 M. J., THOMPSON, C. R., CHAMPION, M. M., ABRAMOVITCH, R. B. &
581 CHAMPION, P. A. 2017. WhiB6 regulation of ESX-1 gene expression is
582 controlled by a negative feedback loop in Mycobacterium marinum. *Proc*
583 *Natl Acad Sci U S A,* 114, E10772-E10781.
584 BOSSERMAN, R. E., NICHOLSON, K. R., CHAMPION, M. M. & CHAMPION, P. A.
585 2019. A New ESX-1 Substrate in Mycobacterium marinum That Is
586 Required for Hemolysis but Not Host Cell Lysis. *J Bacteriol,* 201.
587 BURIAN, J. & THOMPSON, C. J. 2018. Regulatory genes coordinating antibiotic-
588 induced changes in promoter activity and early transcriptional
589 termination of the mycobacterial intrinsic resistance gene whiB7. *Mol*
590 *Microbiol,* 107, 402-415.
591 CHA, R. S., ZARBL, H., KEOHAVONG, P. & THILLY, W. G. 1992. Mismatch
592 amplification mutation assay (MAMA): application to the c-H-ras gene.
593 *PCR Methods Appl,* 2, 14-20.
594 CORTES, T., SCHUBERT, O. T., ROSE, G., ARNVIG, K. B., COMAS, I.,
595 AEBERSOLD, R. & YOUNG, D. B. 2013. Genome-wide mapping of
596 transcriptional start sites defines an extensive leaderless transcriptome
597 in Mycobacterium tuberculosis. *Cell reports,* 5, 1121-31.
598 COUSO, J. P. & PATRAQUIM, P. 2017. Classification and function of small open
599 reading frames. *Nat Rev Mol Cell Biol,* 18, 575-589.
600 COX, J. & MANN, M. 2008. MaxQuant enables high peptide identification rates,
601 individualized p.p.b.-range mass accuracies and proteome-wide protein
602 quantification. *Nat Biotechnol,* 26, 1367-72.
603 CRAPPE, J., VAN CRIEKINGE, W., TROOSKENS, G., HAYAKAWA, E., LUYTEN, W.,
604 BAGGERMAN, G. & MENSCHAERT, G. 2013. Combining in silico prediction
605 and ribosome profiling in a genome-wide search for novel putatively
606 coding sORFs. *BMC Genomics,* 14, 648.
607 DEANA, A. & BELASCO, J. G. 2005. Lost in translation: the influence of
608 ribosomes on bacterial mRNA decay. *Genes Dev,* 19, 2526-33.
609 DEUTSCH, E. W., CSORDAS, A., SUN, Z., JARNUCZAK, A., PEREZ-RIVEROL, Y.,
610 TERNENT, T., CAMPBELL, D. S., BERNAL-LLINARES, M., OKUDA, S.,
611 KAWANO, S., MORITZ, R. L., CARVER, J. J., WANG, M., ISHIHAMA, Y.,
612 BANDEIRA, N., HERMJAKOB, H. & VIZCAINO, J. A. 2017. The
613 ProteomeXchange consortium in 2017: supporting the cultural change in
614 proteomics public data deposition. *Nucleic Acids Res,* 45, D1100-D1106.
615 DUVAL, M., KOREPANOV, A., FUCHSBAUER, O., FECHTER, P., HALLER, A.,
616 FABBRETTI, A., CHOULIER, L., MICURA, R., KLAHOLZ, B. P., ROMBY, P.,
617 SPRINGER, M. & MARZI, S. 2013. Escherichia coli ribosomal protein S1
618 unfolds structured mRNAs onto the ribosome for active translation
619 initiation. *PLoS Biol,* 11, e1001731.
620 EFSTATHIOU, G., ANTONAKIS, A. N., PAVLOPOULOS, G. A., THEODOSIOU, T.,
621 DIVANACH, P., TRUDGIAN, D. C., THOMAS, B., PAPANIKOLAOU, N.,
622 AIVALIOTIS, M., ACUTO, O. & ILIOPOULOS, I. 2017. ProteoSign: an end-

623 user online differential proteomics statistical analysis platform. *Nucleic*
624 *Acids Res,* 45, W300-W306.
625 ELIAS, J. E. & GYGI, S. P. 2007. Target-decoy search strategy for increased
626 confidence in large-scale protein identifications by mass spectrometry.
627 *Nat Methods,* 4, 207-14.
628 FRITH, M. C., FORREST, A. R., NOURBAKHSH, E., PANG, K. C., KAI, C., KAWAI,
629 J., CARNINCI, P., HAYASHIZAKI, Y., BAILEY, T. L. & GRIMMOND, S. M.
630 2006. The abundance of short proteins in the mammalian proteome.
631 *PLoS Genet,* 2, e52.
632 GU, Z., HARROD, R., ROGERS, E. J. & LOVETT, P. S. 1994. Anti-peptidyl
633 transferase leader peptides of attenuation-regulated chloramphenicol-
634 resistance genes. *Proc Natl Acad Sci U S A,* 91, 5612-6.
635 HEMM, M. R., PAUL, B. J., MIRANDA-RIOS, J., ZHANG, A., SOLTANZAD, N. &
636 STORZ, G. 2010. Small stress response proteins in Escherichia coli:
637 proteins missed by classical proteomic studies. *J Bacteriol,* 192, 46-58.
638 HEMM, M. R., PAUL, B. J., SCHNEIDER, T. D., STORZ, G. & RUDD, K. E. 2008.
639 Small membrane proteins found by comparative genomics and ribosome
640 binding site models. *Mol Microbiol,* 70, 1487-501.
641 HENKIN, T. M. & YANOFSKY, C. 2002. Regulation by transcription attenuation in
642 bacteria: how RNA provides instructions for transcription
643 termination/antitermination decisions. *Bioessays,* 24, 700-7.
644 HINNEBUSCH, A. G., IVANOV, I. P. & SONENBERG, N. 2016. Translational control
645 by 5'-untranslated regions of eukaryotic mRNAs. *Science,* 352, 1413-6.
646 HOBBS, E. C., FONTAINE, F., YIN, X. & STORZ, G. 2011. An expanding universe
647 of small proteins. *Current opinion in microbiology,* 14, 167-73.
648 ITO, K. & CHIBA, S. 2013. Arrest peptides: cis-acting modulators of translation.
649 *Annu Rev Biochem,* 82, 171-202.
650 JOHNSTON, H. M., BARNES, W. M., CHUMLEY, F. G., BOSSI, L. & ROTH, J. R.
651 1980. Model for regulation of the histidine operon of Salmonella. *Proc*
652 *Natl Acad Sci U S A,* 77, 508-12.
653 LOI, V. V., ROSSIUS, M. & ANTELMANN, H. 2015. Redox regulation by reversible
654 protein S-thiolation in bacteria. *Front Microbiol,* 6, 187.
655 LOVETT, P. S. & ROGERS, E. J. 1996. Ribosome regulation by the nascent
656 peptide. *Microbiological reviews,* 60, 366-85.
657 MADEIRA, F., PARK, Y. M., LEE, J., BUSO, N., GUR, T., MADHUSOODANAN, N.,
658 BASUTKAR, P., TIVEY, A. R. N., POTTER, S. C., FINN, R. D. & LOPEZ, R.
659 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019.
660 *Nucleic Acids Res,* 47, W636-W641.
661 MARTINI, M. C., ZHOU, Y., SUN, H. & SHELL, S. S. 2019. Defining the
662 Transcriptional and Post-transcriptional Landscapes of Mycobacterium
663 smegmatis in Aerobic Growth and Hypoxia. *Front Microbiol,* 10, 591.
664 MEYDAN, S., MARKS, J., KLEPACKI, D., SHARMA, V., BARANOV, P. V., FIRTH, A.
665 E., MARGUS, T., KEFI, A., VAZQUEZ-LASLOP, N. & MANKIN, A. S. 2019.
666 Retapamulin-Assisted Ribosome Profiling Reveals the Alternative
667 Bacterial Proteome. *Mol Cell,* 74, 481-493 e6.

668 MIRANDA-CASOLUENGO, A. A., STAUNTON, P. M., DINAN, A. M., LOHAN, A. J. &
669 LOFTUS, B. J. 2016. Functional characterization of the Mycobacterium
670 abscessus genome coupled with condition specific transcriptomics
671 reveals conserved molecular strategies for host adaptation and
672 persistence. *BMC Genomics,* 17, 553.
673 NAKAMOTO, T. 2009. Evolution and the universality of the mechanism of
674 initiation of protein synthesis. *Gene,* 432, 1-6.
675 OHNISHI, K., KUTSUKAKE, K., SUZUKI, H. & LINO, T. 1992. A novel
676 transcriptional regulation mechanism in the flagellar regulon of
677 Salmonella typhimurium: an antisigma factor inhibits the activity of the
678 flagellum-specific sigma factor, sigma F. *Mol Microbiol,* 6, 3149-57.
679 OPPENHEIM, D. S. & YANOFSKY, C. 1980. Translational coupling during
680 expression of the tryptophan operon of Escherichia coli. *Genetics,* 95,
681 785-95.
682 PETERS, J. M., VANGELOFF, A. D. & LANDICK, R. 2011. Bacterial transcription
683 terminators: the RNA 3'-end chronicles. *J Mol Biol,* 412, 793-813.
684 RAPPSILBER, J., ISHIHAMA, Y. & MANN, M. 2003. Stop and go extraction tips for
685 matrix-assisted laser desorption/ionization, nanoelectrospray, and
686 LC/MS sample pretreatment in proteomics. *Anal Chem,* 75, 663-70.
687 SBERRO, H., FREMIN, B. J., ZLITNI, S., EDFORS, F., GREENFIELD, N., SNYDER,
688 M. P., PAVLOPOULOS, G. A., KYRPIDES, N. C. & BHATT, A. S. 2019.
689 Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small,
690 Novel Genes. *Cell,* 178, 1245-1259 e14.
691 SHELL, S. S., WANG, J., LAPIERRE, P., MIR, M., CHASE, M. R., PYLE, M. M.,
692 GAWANDE, R., AHMAD, R., SARRACINO, D. A., IOERGER, T. R., FORTUNE,
693 S. M., DERBYSHIRE, K. M., WADE, J. T. & GRAY, T. A. 2015. Leaderless
694 Transcripts and Small Proteins Are Common Features of the
695 Mycobacterial Translational Landscape. *PLoS Genet,* 11, e1005641.
696 SMITH, C., CANESTRARI, J. G., WANG, J., DERBYSHIRE, K. M., GRAY, T. A. &
697 WADE, J. T. 2019. Pervasive Translation in Mycobacterium
698 tuberculosis. *bioRxiv,* 665208.
699 STOJILJKOVIC, I., BAUMLER, A. J. & HANTKE, K. 1994. Fur regulon in gram-
700 negative bacteria. Identification and characterization of new iron-
701 regulated Escherichia coli genes by a fur titration assay. *J Mol Biol,* 236,
702 531-45.
703 SWAMINATHAN, S., ELLIS, H. M., WATERS, L. S., YU, D., LEE, E. C., COURT, D.
704 L. & SHARAN, S. K. 2001. Rapid engineering of bacterial artificial
705 chromosomes using oligonucleotides. *Genesis,* 29, 14-21.
706 TURNBOUGH, C. L., JR. 2019. Regulation of Bacterial Gene Expression by
707 Transcription Attenuation. *Microbiol Mol Biol Rev,* 83.
708 VAN KESSEL, J. C. & HATFULL, G. F. 2008. Efficient point mutagenesis in
709 mycobacteria using single-stranded DNA recombineering:
710 characterization of antimycobacterial drug targets. *Mol Microbiol,* 67,
711 1094-107.

712 WEAVER, J., MOHAMMAD, F., BUSKIRK, A. R. & STORZ, G. 2019. Identifying
713 Small Proteins by Ribosome Profiling with Stalled Initiation Complexes.
714 *MBio,* 10.
715 XU, X., VILCHEZE, C., AV-GAY, Y., GOMEZ-VELASCO, A. & JACOBS, W. R., JR.
716 2011. Precise null deletion mutations of the mycothiol synthesis genes
717 reveal their role in isoniazid and ethionamide resistance in
718 Mycobacterium smegmatis. *Antimicrob Agents Chemother,* 55**,** 3133-9.
719 YANOFSKY, C. 1981. Attenuation in the control of expression of bacterial
720 operons. *Nature,* 289**,** 751-8.
721 ZHAO, G., WEATHERSPOON, N., KONG, W., CURTISS, R., 3RD & SHI, Y. 2008. A
722 dual-signal regulatory circuit activates transcription of a set of divergent
723 operons in Salmonella typhimurium. *Proc Natl Acad Sci U S A,* 105**,**
724 20924-9.
725 ZHENG, X., HU, G. Q., SHE, Z. S. & ZHU, H. 2011. Leaderless genes in bacteria:
726 clue to the evolution of translation initiation mechanisms in prokaryotes.
727 *BMC genomics,* 12**,** 361.
728 ZOUGMAN, A., SELBY, P. J. & BANKS, R. E. 2014. Suspension trapping (STrap)
729 sample preparation method for bottom-up proteomics analysis.
730 *Proteomics,* 14**,** 1006-0.
731

732

733 FIGURE LEGENDS

734 Figure 1. Schematic comparison of structural features of operon attenuation. A. In
735 transcriptional attenuation, the balance of dueling mRNA hairpin structures is controlled
736 by ribosome occupancy of the sensory sORF. Rapid ribosome transit--when $tRNA^{trp}$ is
737 plentiful, for example--allows the vacated sORF mRNA to hybridize with
738 complementary sequence thereby forming the stem of an intrinsic transcriptional
739 terminator. This transcriptional termination keeps the downstream tryptophan
740 biosynthetic genes from unneeded expression. Under conditions of reduced $tRNA^{trp}$
741 levels, ribosome pausing in the sORF blocks the attenuating transcriptional terminator
742 from forming, an anti-terminator structure forms instead, which promotes RNA
743 polymerase extension into the operon. Transcriptional attenuation through controlled
744 intrinsic termination occurs in *E.coli* tryptophan and *S. typhimurium* histidine
745 biosynthetic loci, and the whiB7 locus in *Mycobacterium abscessus* (Burian and
746 Thompson, 2018, Johnston et al., 1980, Yanofsky, 1981). B. In translational attenuation,
747 sensory sORF mRNA base pairing obscures the SD of the downstream gene, preventing
748 translation initiation. Ribosome arrest in the sORF prevents duplex formation, and
749 consequently frees the SD to recruit and position translation initiation complexes for
750 expression of the annotated gene. Translation attenuation has been described for some
751 ribosome antibiotics (Arenz et al., 2014, Lovett and Rogers, 1996, Ito and Chiba, 2013).
752

753 Figure 2. Polycysteines are enriched in LL-sORFs. The frequency of consecutive amino
754 acids is shown in annotated genes (X-axis) and in LL-sORFs (Y-axis). Frequencies for

755 each clustered amino acid (at least two consecutive) are expressed as percentages of its
756 total in each group (see Table S2 for tallies).
757
758 Figure 3. The architecture of the *Ms5788A* to *Ms5790* locus and the polycysteine-
759 encoding mRNA leader. The amino acid sequence of the wild-type Ms5788A includes 9
760 cysteine codons (bold C), with eight consecutive codons at the C-terminus. This leader
761 sequence to *Ms5788* is predicted to fold into a stable duplex structure (Fig S2),
762 represented here in schematic form and annotated with features relevant for regulated
763 expression of *Ms5788*. The ORF defined by the initiating GUG fMet is shown as a
764 magenta line, with key sequences indicated.
765
766 Figure 4. Hypothesized model and ribosome dependence of *Ms5788A*-directed
767 attenuation. A. *Ms5788A* is transcribed (RNA polymerase, RNAP) and translated (orange
768 ribosome). Ribosomes waiting (left) for charged $tRNA^{cys}$ prevent the leader from folding
769 into a stable, attenuating, conformation (right). The nascent small protein is shown
770 emerging from the elongating ribosome and is annotated with the initiating methionine
771 (fM) and polycysteine (C) tract for reference. Except where noted, luciferase reporters are
772 translational fusions of the NanoLuc ORF in-frame with the AUG initiating methionine
773 of the annotated *Ms5788* gene. B. The same wild-type (i) reporter cysteine responsive
774 data are shown in Figs 4-6 (black outlined or filled boxes) as a comparative reference for
775 the mutant derivatives of each series (blue outlined or filled boxes in Fig 4). Non-start (ii)
776 and nonsense (iii) mutations reduce ribosome occupancy and increase attenuation of the
777 reporter. An out-of-frame mutation permits ribosome occupancy but is not cysteine
778 responsive (iv). Lower case letters below clones indicate nucleotide substitutions and

779 insertion/deletions (+/-) used to create mutants. Luciferase activity is shown as light units
780 (x 10^9) per ml of cell culture, and asterisks indicate significance p <.01 for cysteine
781 supplementation, by two-tailed t-test in Figs 4-6.
782
783 Figure 5. Predicted duplex mRNA interactions are required for the attenuated state of the
784 *Ms5788* luciferase reporter fusion. Clustered point mutations were introduced into the
785 *Ms5788A* LL-sORF or the predicted pairing nucleotides to disrupt base-pairing but, when
786 combined, should restore the modeled stable mRNA structure. Two series of mutants
787 were generated (angled hatch fill, ii – iv and horizontal and vertical hatch fill, v – vii).
788 The effect of structure-destabilizing mutations and a re-stabilizing combination were
789 tested for their effect on the luciferase reporter fused to *Ms5788.* All three clustered
790 mutants (ii, iii, v) that are expected to disrupt the predicted duplex structure model an
791 unattenuated state, whereas a G:U stabilized mutant (vi) or combined complement
792 mutants (iv, vii) retain cysteine responsiveness. See Fig S3 for G:U stabilized structure
793 detail.
794
795 Figure 6. Attenuation regulates translation initiation of *Ms5788*. Fifteen nt were inserted
796 at the end of the wild-type leader and included an independent SD to assess mRNA
797 extension beyond the structured attenuator (ii). Luciferase activity is now independent of
798 the effects of cysteine supplementation on the attenuating structure. Placement of the
799 luciferase reporter at the initiation codon of *Ms5789* (iii) indicates that translation
800 attenuation in *Ms5788* has polar effects, propagating the control of *Ms5788A* to operonic
801 genes.

802 Figure 7. Cysteine attenuation is observed in the native context. A. Label-Free
803 Quantitative proteomics (LFQ) was used to identify and quantitate changes in the
804 abundance of proteins from whole-cell extracts of wild-type *M. smegmatis* subjected to
805 trypsin digestion and nanoUHPLC-MS/MS. *M. smegmatis* was cultured in minimal
806 medium (X-axis) or minimal medium supplemented with cysteine (Y-axis). Values
807 plotted are normalized LFQ peak area (protein) in arbitrary units. The baselines for both
808 X and Y-axes were artificially set at 5 x10^4 counts (indicated by dashed limit of
809 detection) to allow depiction of proteins not expressed in one of the two conditions.
810 Peptides from Ms5789 and Ms5790 (white diamond 1 and 2, respectively) increased in
811 abundance (shift right along the X-axis) under cysteine limitation. Peptides of annotated
812 proteins from loci similarly encoded on polycysteine-encoding LL-sORF mRNAs are
813 indicated by black filled circles (Fig S4; 3 = Ms0113; 4 = Ms0114; 5 = Ms0934; 6 =
814 Ms4527; 7 = Ms4533; 8 = Ms5279; 9 = Ms5280). B. A frameshift mutation (OoF) in
815 *Ms5788A* that changes the CCCCCCCC* to VAVVVVAVERSRAL*. This OoF LL-
816 sORF mutant is insensitive to cysteine and does not release Ms5789 and Ms5790 from an
817 attenuated state. C. Deleting *Ms5788A* elevates expression of Ms5789 and Ms5790 and is
818 insensitive to cysteine, indicating a completely unattenuated state.
819
820 Figure 8. Phylogenetic distribution of polycysteine LL-sORFs in mycobacteria. A robust
821 reference phylogenetic tree was constructed from complete genome sequences of 41
822 diverse *Mycobacterium* spp. All nodes had bootstrap values of 100, except where
823 indicated. Species of the slow-grower clade appear in blue text, and fast-growers in
824 magenta. The presence (black square) or absence (white square) of each LL-sORF forms

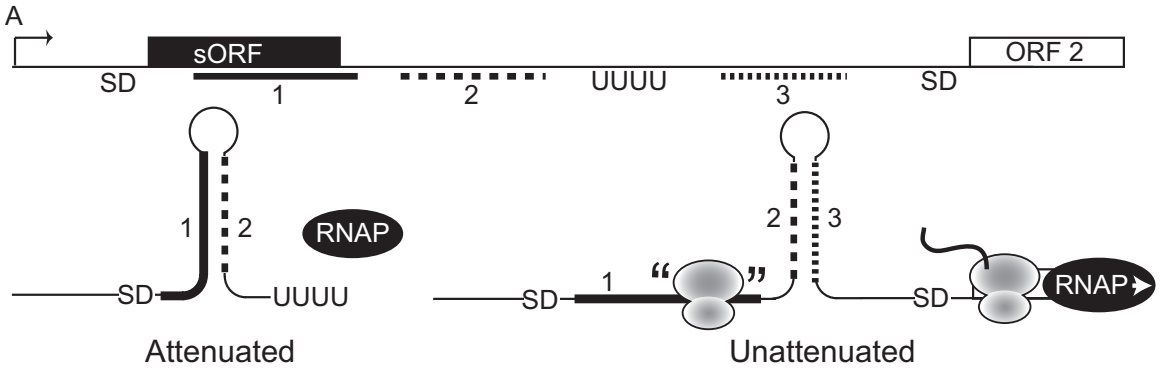
825 a binary barcode for each species. For example, the presence of *Ms0113A* in some fast-
826 growing species is indicated by a black square in the leftmost box of the barcode.
827 Barcode key and fixed-length sequence logo for each LL-sORF is shown at right.
828
829 Figure 9. Schematic comparison of a traditional transcriptionally coordinated regulon and
830 the translationally controlled ribulon. A. Regulons are usually controlled at the
831 transcriptional level and are named for their master regulator (MR). In this scenario,
832 coordinate gene expression from multiple loci, is mediated by the MR activating
833 transcription. The MR is shown binding to multiple loci, which activates transcription
834 and subsequent translation by the ribosome. This includes activation of sigma factors
835 (e.g., Sigma F flagellar regulon in *Salmonella typhimurium* (Ohnishi et al., 1992)),
836 inactivation of a repressor (e.g., *fur* regulon in *E. coli* (Stojiljkovic et al., 1994)), or two-
837 component activation (e.g., *slyA* regulon in *Salmonella typhimurium* (Zhao et al., 2008)).
838 B. In the Cys ribulon, global gene coordination occurs by mutual pausing of ribosomes in
839 ribulon sORFs. Attenuation at individual LL-sORFs at independent operons is co-
840 dependent on $tRNA^{cys}$ levels, resulting in coordinated gene expression. In cysteine replete
841 conditions, polycysteine sORFs are rapidly translated and expression of downstream
842 genes is repressed. Under conditions of limiting $tRNA^{cys}$ levels, ribosomes pause at
843 polycysteine tracts in the sensory LL-sORFs and relieve attenuation of operonic genes to
844 upregulate processes required for the production of cysteine. Sensory sORFs are shown
845 as black boxes, stalled ribosomes are red-shaded, activated genes are indicated by open
846 rectangles and translated by elongating ribosomes (green-shaded) with the emerging
847 nascent protein.

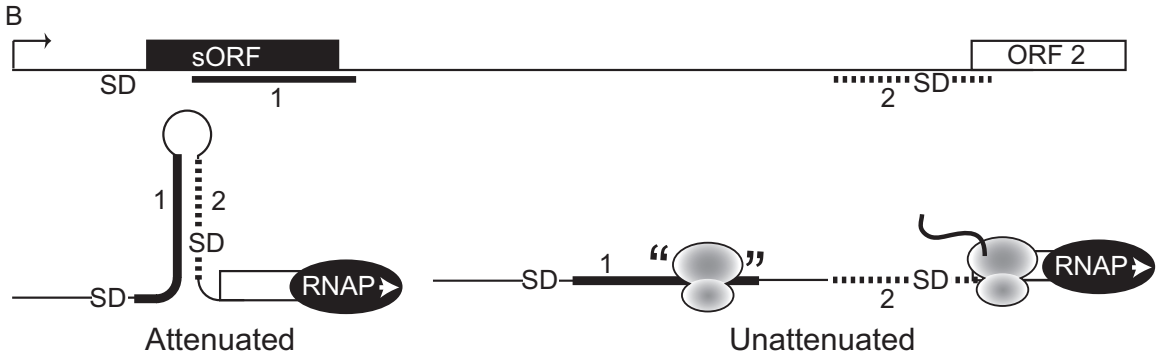
gene #	name	encoded protein predicted function	HMM Pfam	uniprot
Ms0113A		putative attenuating leaderless polycysteine sORF (this report)		
Ms0113	tauC	taurine transport system permease protein TauC; sulfur starvation	PF00528	A0QNP1
Ms0114		hnD/SsuA/transferrin family substrate-binding protein	PF04069	A0QNP2
Ms0116		taurine import ATP-binding protein TauB	PF00005	A0QNP3
Ms0932A		putative attenuating leaderless polycysteine sORF (this report)		
Ms0932		ROK family protein	PF00480	A0QQZ7
Ms0933	mshA	MshA glycosyltransferase	PF00534	A0QQZ8
Ms0934		YbjN domain-containing protein		A0QQZ9
Ms4527A		putative attenuating leaderless polycysteine sORF (this report)		
Ms4527		ferredoxin sulfite reductase	PF01077	A0R0W1
Ms4528	cysH	phosphoadenosine phosphosulfate reductase	PF01507	A0R0W2
Ms4529		secreted protein, sirohydrochlorin chelatase	PF01903	A0R0W3
Ms4530	cysA	sulfate ABC transporter, ATP-binding protein	PF00005	A0R0W4
Ms4533A		putative attenuating leaderless polycysteine sORF (this report)		
Ms4533		sulfate ABC transporter substrate-binding protein	TIGR00971	A0R0W7
Ms4532	cysT	sulfate ABC transporter, permease protein CysT	PF00528	A0R0W6
Ms4531	cysW	sulfate ABC transporter, permease protein	PF00528	A0R0W5
Ms4536A		putative attenuating leaderless polycysteine sORF (this report)		
Ms4536		TQXA domain-containing protein; surface-expressed		A0R0X0
Ms5280A		putative attenuating leaderless sORF (this report)		
Ms5280		cysteine dioxygenase type I superfamily protein	PF05995	A0R2Y8
Ms5279		sulfurtransferase		A0R2Y7
Ms5788A		putative attenuating leaderless polycysteine sORF (this report)		
Ms5788		DUF4395 domain-containing protein		A0R4C8
Ms5789	cysA2	putative thiosulfate sulfurtransferase	PF00581	A0R4C9
Ms5790		DUF1416 domain-containing protein, SseC protein	PF07210	A0R4D0
Rv2334A		putative attenuating leaderless polycysteine sORF (this report)		
Rv2334	cysK1	O-acetylserine sulfhydrylase; Interacts with CysE		P9WP55
Rv2335	cysE	serine acetyltransferase; acetylates serine to make cysteine		P95231

Table 1. Cysteine translation attenuation regulon gene annotations. Gene order, content, and identifiers are listed each putative regulon operon for *M. smegmatis* (Ms). Only the *Rv2334A* operon is shown for *M. tuberculosis* (Rv) . Operons are alternately boxed for clarity of separation. The predicted functions and features of the encoded proteins are shown (Pfam or uniprot).

857 FIGURES

858 Fig 1





859

860

861 Fig 2

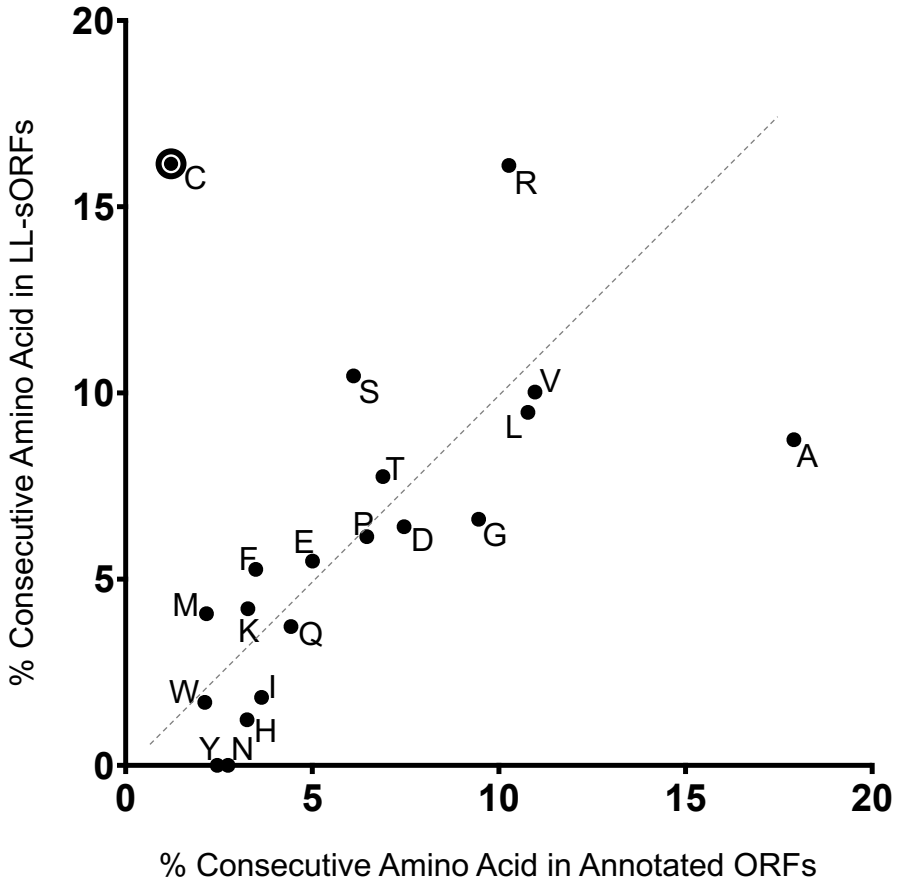

% Consecutive Amino Acid in LL-sORFs
% Consecutive Amino Acid in Annotated ORFs
0
5
10
15
20
C
R
S
V
L
A
T
P
D
G
E
F
M
K
Q
W
I
H
Y
N

Fig 3

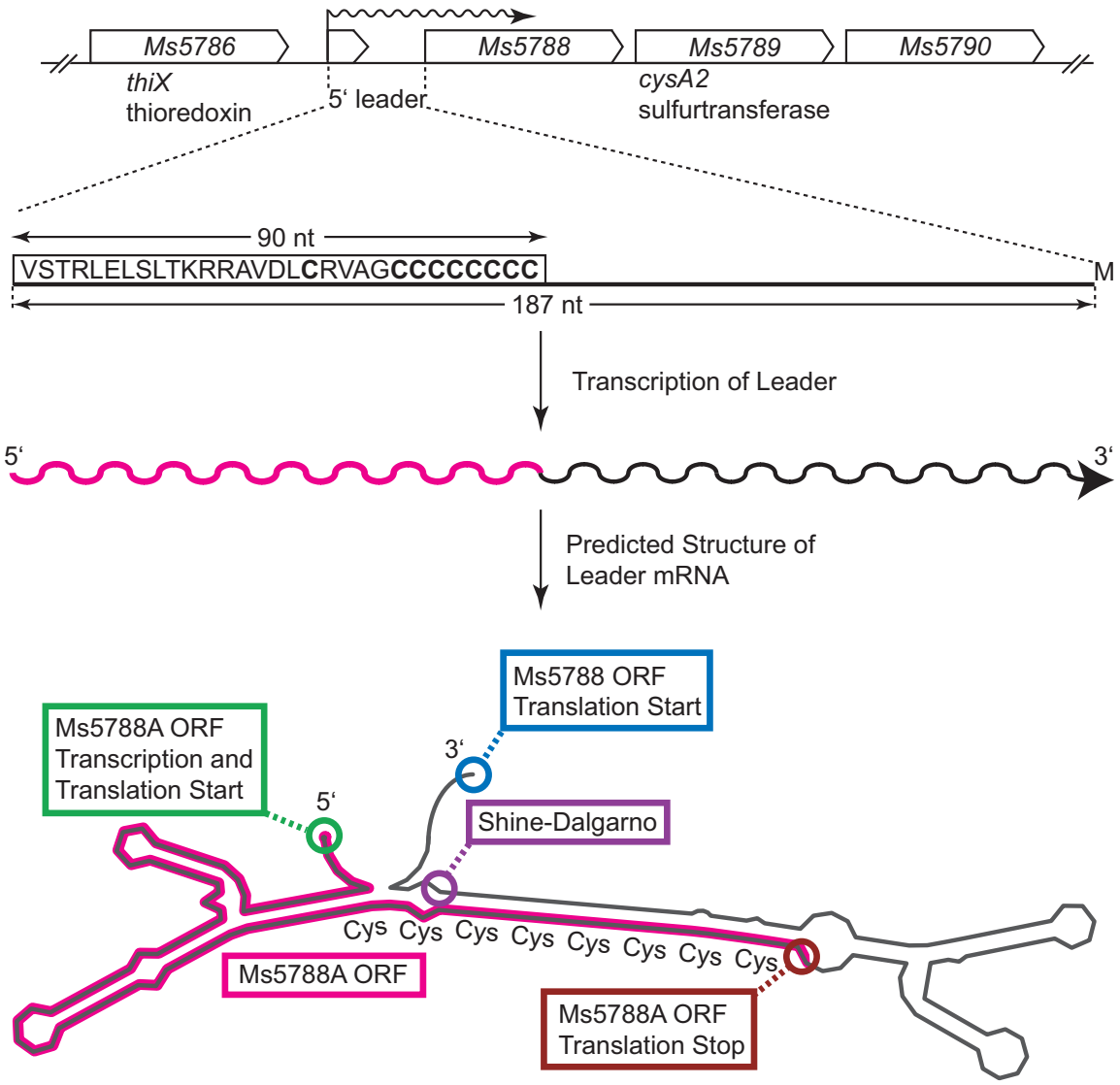
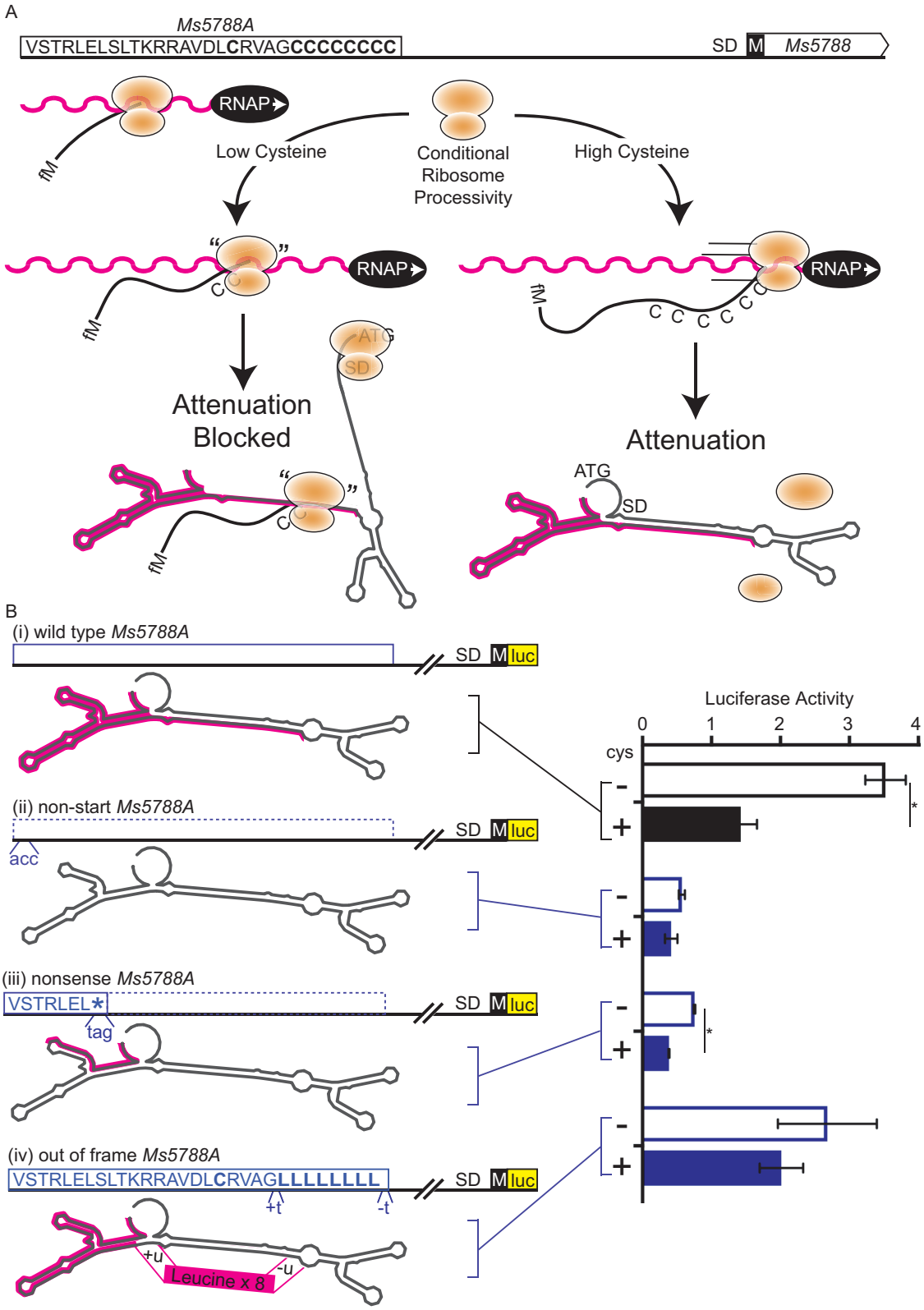

Ms5786
Ms5788
Ms5789
Ms5790
thiX
thioredoxin
5‘ leader
cysA2
sulfurtransferase
90 nt
VSTRLELSLTKRRAVDLCRVAGCCCCCCCC
M
187 nt
Transcription of Leader
5‘
3‘
Predicted Structure of
Leader mRNA
Ms5788 ORF
Translation Start
Ms5788A ORF
Transcription and
Translation Start
3‘
5‘
Shine-Dalgarno
Cys Cys Cys Cys Cys Cys Cys Cys
Ms5788A ORF
Ms5788A ORF
Translation Stop

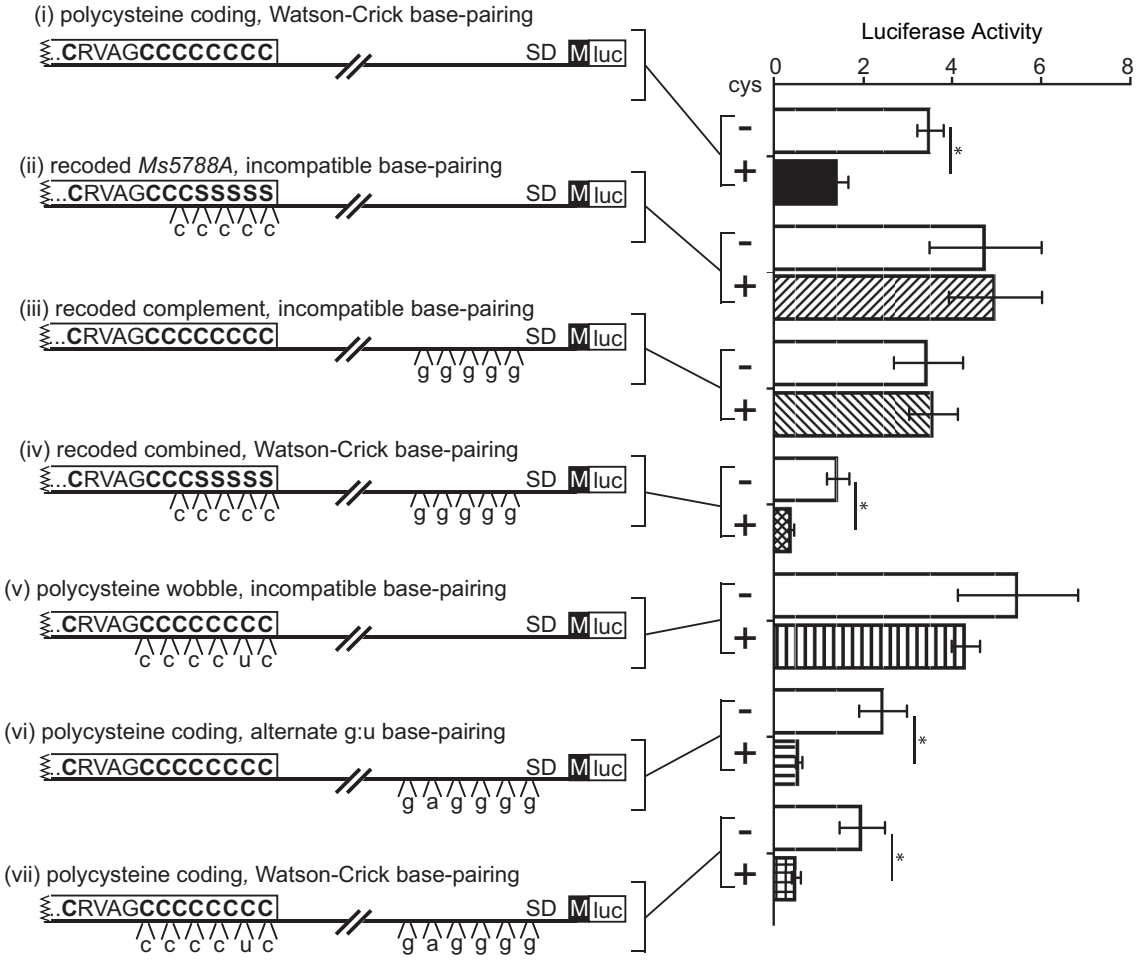
Fig 4


A
Ms5788A
VSTRLELSLTKRRAVDLCRVAGCCCCCCCC
SD M Ms5788
RNAP
fM
Low Cysteine
Conditional Ribosome Processivity
High Cysteine
C
C C C C C C
ATG
SD
Attenuation Blocked
Attenuation
B
(i) wild type Ms5788A
SD M luc
Luciferase Activity
0 1 2 3 4
cys
-
+
*
(ii) non-start Ms5788A
acc
(iii) nonsense Ms5788A
VSTRLEL*
tag
(iv) out of frame Ms5788A
VSTRLELSLTKRRAVDLCRVAGLLLLLLLL
+t
-t
+u
-u
Leucine x 8

868

869

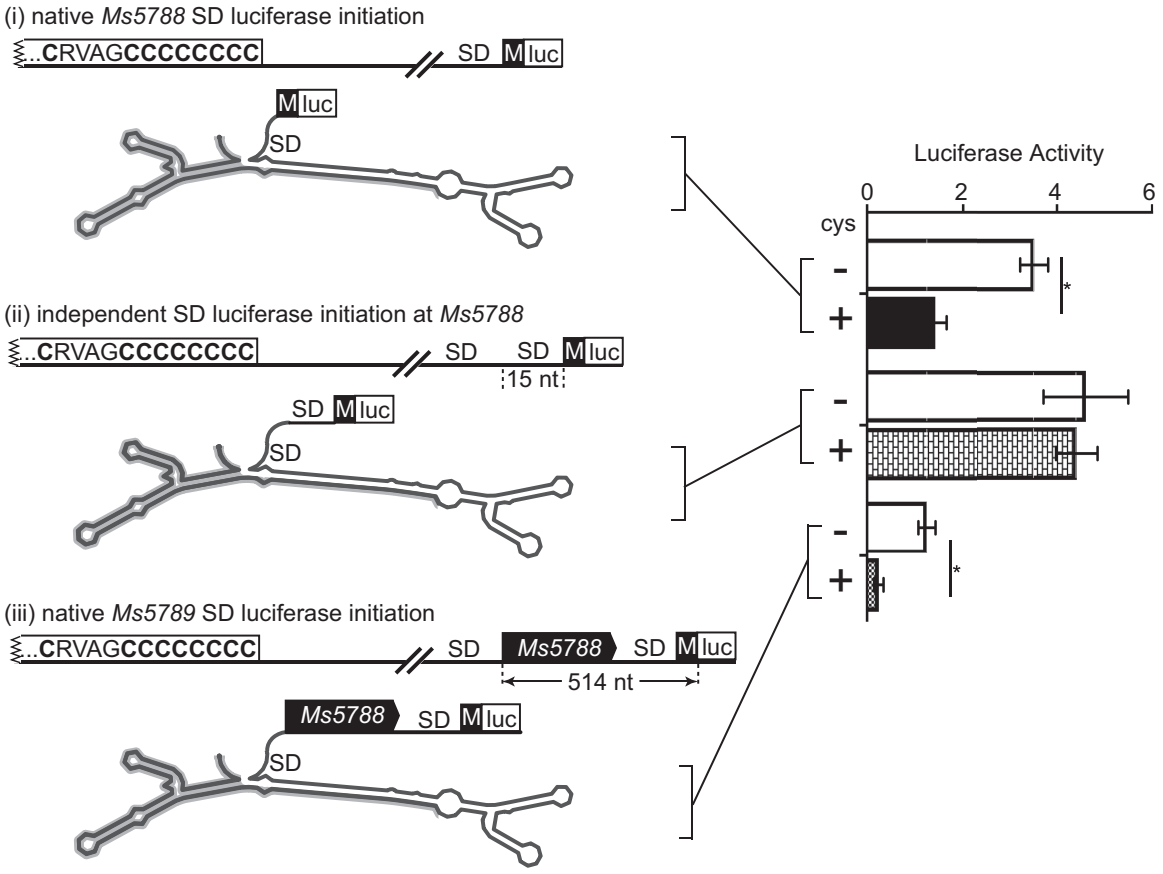
870 Fig 5



871

872

873 Fig 6



874

875

876 Fig 7

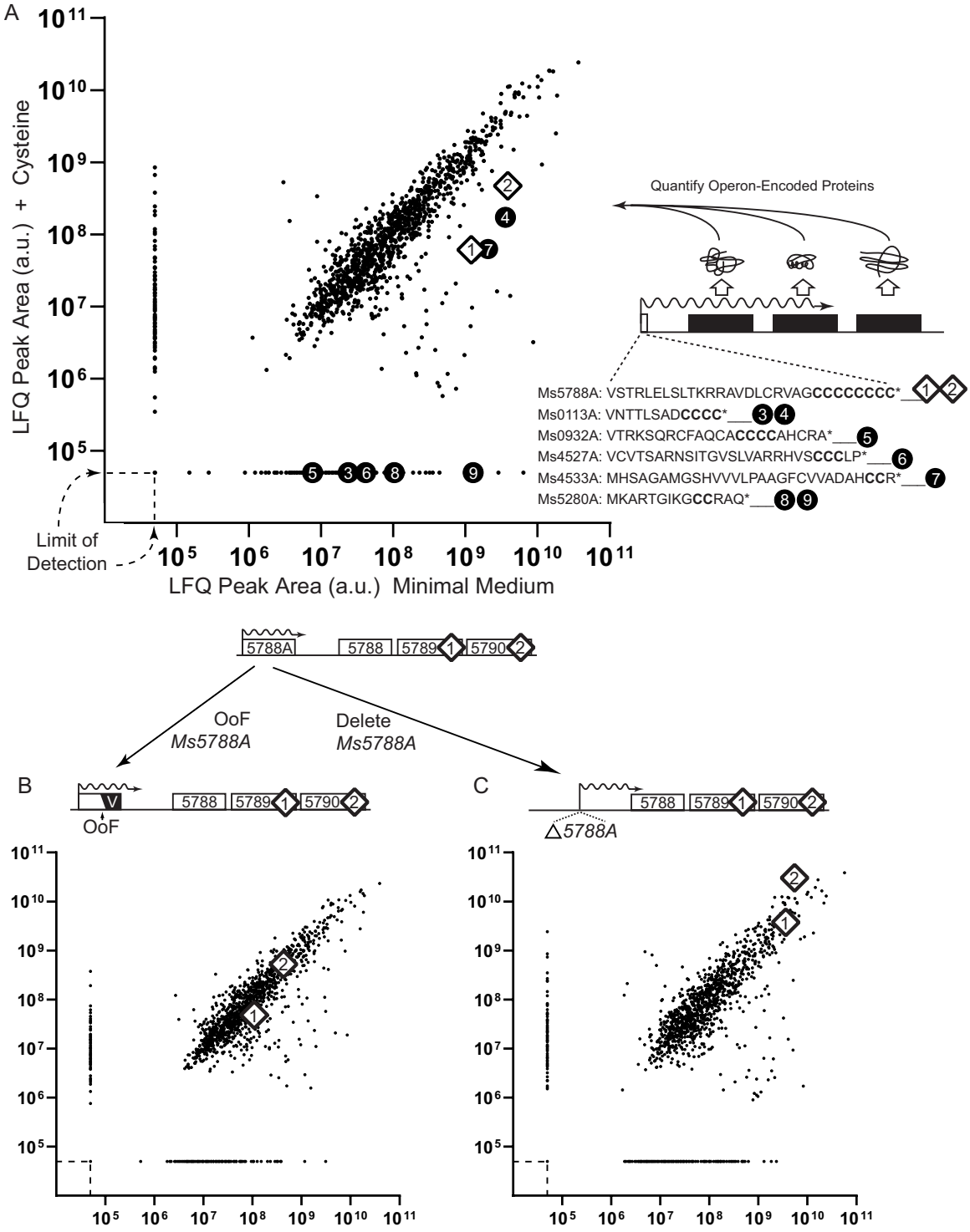
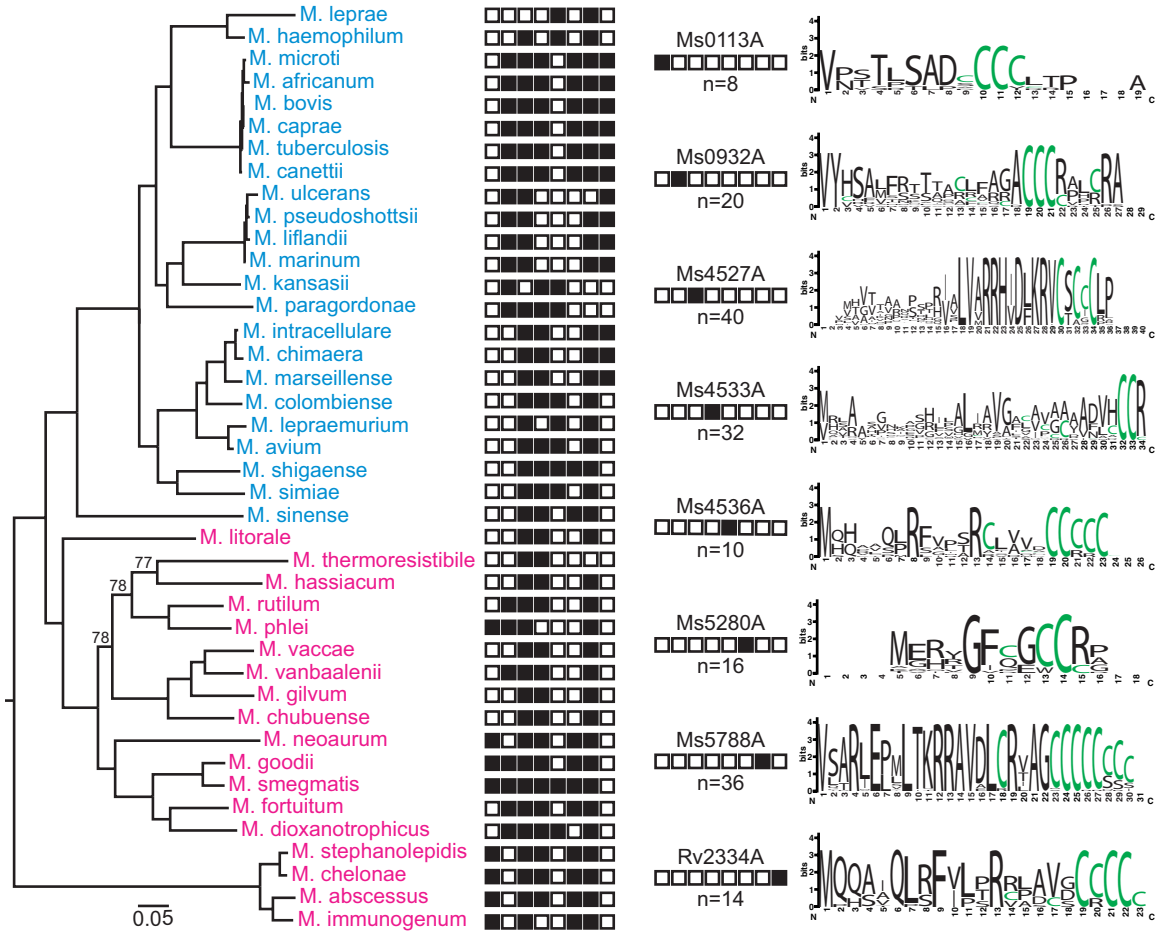

A
LFQ Peak Area (a.u.) + Cysteine
LFQ Peak Area (a.u.) Minimal Medium
Limit of Detection
Quantify Operon-Encoded Proteins
Ms5788A: VSTRLELSLTKRRAVDLCRVAGCCCCCCCC*
Ms0113A: VNTTLSADCCCC*
Ms0932A: VTRKSQRCFAQCACCCCAHCRA*
Ms4527A: VCVTSARNSITGVSLVARRHVSCCCLP*
Ms4533A: MHSAGAMGSHVVVLPAAGFCVVADAHCCR*
Ms5280A: MKARTGIKGCCRAQ*
5788A
5788
5789
5790
OoF
Ms5788A
Delete
Ms5788A
B
OoF
C
Δ5788A

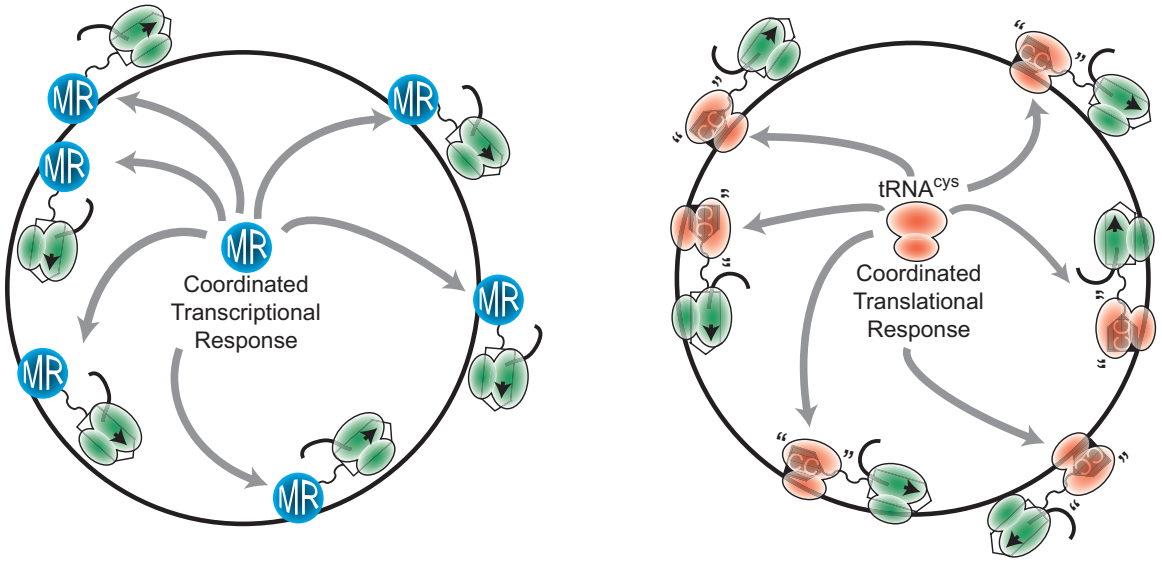
Fig 8



880

881

882 Fig 9



883