Leaderless short ORFs in mycobacteria comprise a translational 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
Attenuation
Cysteine
Mycobacteria
Leaderless mRNAs
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, meaning the RNAs begin with a 5'-AUG or GUG initiation codon. Leaderless mRNAs often encode an unannotated short ORF as the first gene of a polycistronic transcript. Consecutive cysteine codons are highly overrepresented in mycobacterial leaderless short ORFs. Here we show that polycysteine-encoding leaderless short ORFs function as cysteine-responsive attenuators of operonic gene expression. Through detailed mutational analysis, we show that one such polycysteine-encoding short ORF controls expression of the downstream genes by causing ribosome stalling under conditions of low cysteine. Ribosome stalling in turn blocks mRNA secondary structures that otherwise sequester the Shine-Dalgarno ribosome-binding site of the 3'gene. This translational attenuation does not require competing transcriptional terminator formation, a mechanism that underlies traditional amino acid attenuation systems. We further assessed cysteine attenuation in *Mycobacterium smegmatis* using mass spectrometry to evaluate endogenous proteomic responses. Notably, six cysteine metabolic loci that have unannotated polycisteine-encoding leaderless short ORF architectures responded to cysteine supplementation/limitation, indicating that cysteine-responsive attenuation is widespread in mycobacteria. Individual leaderless short ORFs confer independent operon-level control, while their shared dependence on

cysteine ensures a collective response. Bottom-up regulon coordination is the antithesis of traditional top-down master regulator regulons and illustrates one utility of the many unnanotated short ORFs expressed in bacterial genomes.

51 INTRODUCTION

52

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

72

73 Many ORFs initiated by leaderless mRNAs (LL-mRNAs) in mycobacteria are short
74 (defined here as less than 150 nt) and unannotated. Importantly, the ORFs initiated at the
75 combined transcription/translation initiation start sites of LL-mRNAs are readily
76 identifiable from transcriptomic data sets and thus provide a high-confidence list of
77 expressed novel mycobacterial sORFs. Insights into the mechanistic and functional
78 attributes of LL-mRNAs have lagged, as they are rare or poorly expressed in *E. coli* but
79 are abundant in archaea, Actinobacteria and extremophiles (Beck and Moll 2018). LL-
80 sORFs represent the first (5'-most) ORF of a transcript, which positions them for a role in
81 *cis*-regulation of the downstream operonic genes. There is ample precedent in eukaryotes
82 for regulation of downstream genes by such "upstream ORFs" (uORFs) (Hinnebusch et
83 al. 2016; Couso and Patraquim 2017). In prokaryotes, mechanisms have also been
84 previously described in which uORFs have been shown to regulate expression of
85 downstream genes through a process known as attenuation (Oppenheim and Yanofsky
86 1980; Bechhofer 1990; Henkin and Yanofsky 2002).
87
88 Attenuation is a *cis*-regulatory mechanism often mediated by short uORFs enriched in
89 codons for the amino acid product of that biosynthetic operon. Attenuation occurs when
90 abundant charged tRNA levels allow translating ribosomes to quickly clear the
91 modulating uORF, promoting the formation of an intrinsic terminator that aborts
92 transcription of the operon. Low levels of charged tRNA cause ribosome stalling in the
93 uORF at codons for the end-product amino acid, facilitating the formation of a competing
94 anti-terminator structure, thereby releasing attenuation to allow transcription to extend
95 into the biosynthetic operon (Turnbough 2019). uORF-mediated attenuation mechanisms

96 for cysteine have not been described. A subset of predicted mycobacterial LL-sORFs
97 conspicuously encode consecutive cysteine residues, and these were found upstream of
98 genes annotated to be involved in cysteine biosynthesis (Shell et al. 2015). We
99 hypothesized that these LL-sORFs function as cysteine-sensitive attenuators.
100
101 RESULTS
102
103 We identified 304 putative LL-sORFs in *Mycobacterium smegmatis* (Supplementary
104 Information Table 1(Shell et al. 2015; Martini et al. 2019)). We compared amino acid
105 content of the encoded sproteins and found that consecutive cysteines were
106 overrepresented relative to cysteine content (chi-square $p<.01$, Extended Data Table 1)
107 and relative to consecutive cysteine frequency in annotated genes. A subset of predicted
108 mycobacterial LL-sORFs conspicuously encode consecutive cysteine residues, and these
109 were often found upstream of genes annotated to be involved in cysteine biosynthesis
110 (Shell et al. 2015). We hypothesized that these LL-sORFs function as cysteine-sensitive
111 attenuators.
112
113 ***Ms5788* is regulated in response to cysteine abundance**
114 One predicted LL-sORF (here denoted Ms5788A) encodes eight consecutive cysteines in
115 its C-terminus, and is followed by operonic genes, including a putative thiosulfate
116 sulfurtransferase, *cysA2* (Fig 1A). RNA-seq and Ribo-seq profiles indicate that this
117 unannotated LL-sORF is abundantly transcribed and translated in *M. smegmatis*
118 (Extended Data Fig 1 A). Moreover, transcription start site mapping strongly suggests

119 that a homologous LL-sORF is expressed in *M. tuberculosis* (Extended Data Fig 1B). To
120 determine whether the genes located immediately downstream of *Ms5788A* are regulated
121 in response to changes in cysteine levels, we generated a luciferase translational reporter
122 in which a constitutive promoter drives a leaderless transcript that begins at the native
123 GUG initiation codon of *Ms5788A* and continues to the initiation codon of the annotated
124 gene downstream, *Ms5788* (Fig 1B). We then measured expression of the reporter in *M.*
125 *smegmatis* cells in the presence or absence of cysteine in the growth medium. Luciferase
126 activity decreased for cells grown with cysteine supplementation (Fig 1B i). Hence, we
127 hypothesized that expression of *Ms5788* is regulated in response to cysteine levels by an
128 attenuation mechanism involving the upstream LL-sORF *Ms5788A*.
129
130 We next tested whether cysteine-dependent regulation of *Ms5788* requires translation of
131 the upstream *Ms5788A* LL-sORF. We mutated the GUG translation initiation codon of
132 *Ms5788A* to ACC in the context of the luciferase reporter, to prevent ribosome loading.
133 This non-start mutation reduced luciferase expression, and abolished attenuation (Fig 1B
134 compare i vs ii). These data indicate that translation of *Ms5788A* is required for cysteine-
135 dependent regulation of *Ms5788*, and that in the absence of *Ms5788A* translation,
136 expression of *Ms5788* is locked in an attenuated state. We speculated that ribosome
137 occupancy of the C-terminal polycysteine tract of *Ms5788A* is particularly important for
138 this regulation. We created a nonsense mutant, Ser8Stop, to truncate *Ms5788A* ten amino
139 acids before the first Cys codon. This truncating mutation also dramatically reduced
140 luciferase expression and attenuation by cysteine (Fig 1B iii). The residual cysteine

141 response of this nonsense mutant may result from translation reinitiation after the stop
142 codon.
143
144 Since ribosome occupancy of the LL-sORF appeared to be required for cysteine-
145 dependent regulation of the luciferase reporter, we postulated that ribosomes stalled in
146 the *Ms5788A* polycysteine tract due to limiting levels of charged $tRNA^{cys}$ would increase
147 luciferase expression by relieving attenuation. We hypothesized that recoding the
148 polycysteine tract should impair the observed cysteine response. We created an out-of-
149 frame (OoF) *Ms5788A* mutant luciferase reporter to replace the eight consecutive
150 cysteine codons with eight consecutive leucine codons, leaving a single Cys18 codon
151 (Fig 1B iv). This OoF mutant was not affected by cysteine supplementation, indicating
152 the importance of the polycysteine tract in relieving attenuation. Interestingly, expression
153 of this reporter appears to be in the active state, suggesting that limiting leucine may
154 functionally substitute for limiting cysteine.
155
156 **Ribosome stalling in the *Ms5788A* sORF modulates RNA structure in the *Ms5788* 5'**
157 **UTR**
158 Ribosome occupancy of *Ms5788A* could affect the formation of mRNA secondary
159 structures in the *Ms5788* mRNA leader, as occurs in previously described attenuation
160 mechanisms (Turnbough 2019). The predicted RNA secondary structure for the mRNA
161 through the *Ms5788A* LL-sORF and up to *Ms5788*, indicates the potential for an
162 energetically stable structure over most of its length (Fig 2). Importantly, nucleotides in
163 the polycysteine tract of *Ms5788A* are predicted to base pair with complementary

164 sequences near the Shine-Dalgarno sequence of *Ms5788*. This structure suggests a
165 mechanism in which stalled ribosomes in *Ms5788A* free the Shine-Dalgarno sequence to
166 recruit and position ribosomes for canonical translation initiation of *Ms5788*.
167
168 Whereas the *Ms5788A* OoF mutant was predicted to have no effect on mRNA structure,
169 we created mutants intended to selectively disrupt predicted duplex pairing near the
170 *Ms5788* Shine-Dalgarno sequence (Fig 2). We first changed the invariant guanine in the
171 2nd position of cysteine codons (U<u>G</u>Y) to cytosine (U<u>C</u>Y) in the last five codons of
172 *Ms5788A* (Fig 2, recoded red series bottom strand). The nucleotide changes should
173 reduce the stability of base pair interactions with the Shine-Dalgarno region, while
174 recoding polyserine for the final five codons of *Ms5788A*: Cys(26-30)Ser. This reporter
175 exhibited constitutively elevated expression that was insensitive to cysteine (Fig 2 ii),
176 consistent with full Shine-Dalgarno sequence accessibility.
177
178 To differentiate the effect of RNA duplex formation from the effect of *Ms5788A* amino
179 acid recoding, we created a mutant that only disrupted the base pairing, by changing the
180 predicted five cognate nucleotides near the Shine-Dalgarno sequence (Fig 2, recoded red
181 top strand). Even though the polycysteine tract remained intact, luciferase expression
182 from this reporter was insensitive to cysteine supplementation (Fig 2 iii), consistent with
183 the base-paired structure corresponding to the attenuated state. We next constructed a
184 mutant that combined the recoded *Ms5788A* and cognate non-coding mutants, such that
185 base pairing should be restored. As expected, combining the complementary mutations
186 reduced expression of the luciferase reporter for cells grown in the presence of cysteine,

187 consistent with restored base pairing (Fig 2 iv). Interestingly, the restored response to
188 cysteine indicated that the residual four-cysteine codon content of the LL-sORF was
189 sufficient to confer sensitivity to cysteine.
190
191 We constructed a silent *Ms5788A* mutant that switched the nucleotide sequence of six
192 cysteine codons to retain polycysteine coding, and yet are predicted to disrupt base
193 pairing with the Shine-Dalgarno region (Fig 2 polycysteine purple bottom strand). The
194 elevated luciferase activity of this mutant reporter was not attenuated under cysteine-
195 replete conditions, separating the roles of nucleotide and encoded amino acid sequence of
196 *Ms5788A* (Fig 2 v). We also created a mutant of the predicted complementary bases near
197 the Shine-Dalgarno sequence (Fig 2, polycysteine purple top strand). Surprisingly, this
198 mutant exhibited an attenuation response similar to wild type (Fig 2, compare i and vi).
199 We reassessed the base pairing potential of the mRNA produced from this mutant and
200 found that it was predicted to fold into a stable, wild-type-like structure that would also
201 reduce Shine-Dalgarno sequence availability (Extended Data Fig 2). Combining the
202 *Ms5788A* and peri-Shine-Dalgarno nucleotide changes in this series resulted in an
203 expression pattern similar to that of the wild-type construct, consistent with the model
204 (Fig 2 vii).
205
206 **Cysteine-dependent regulation involving *Ms5788A* affects expression of downstream**
207 **operonic genes**
208 Classical models of ribosome-mediated attenuation invoke competing mRNA stem loops
209 that form an intrinsic terminator structure when ribosomes rapidly translate the sORF,

210 resulting in regulation at the level of transcription (Yanofsky 1981; Turnbough 2019). In
211 the case of *Ms5788* attenuation, regulation appears to occur at the level of translation. To
212 test this, we created a reporter to assess mRNA extension beyond the start of *Ms5788*. To
213 maintain the predicted RNA structure of the 5' leader region, an independent Shine-
214 Dalgarno sequence was added to efficiently initiate translation of transcripts that extend
215 into the luciferase gene (Fig 3 ii). Luciferase activity of this reporter was insensitive to
216 cysteine, and expression levels were consistently high. These data indicate that *Ms5788*
217 attenuation does not result from transcription termination. Translational repression can
218 indirectly affect transcription over longer distances due to polarity that is caused by Rho-
219 dependent transcription termination and/or by enhanced RNase processing of the
220 untranslated RNA (Deana and Belasco 2005; Martini et al. 2019). To determine if
221 translational repression of *Ms5788* by attenuation leads to polar effects on downstream
222 genes, we constructed a translational fusion that extends to *Ms5789* (*cysA2*). This more
223 distal site, 516 nt 3' of the *Ms5788* start, exhibited cysteine responsiveness (Fig 3 iii).
224 Taken together, our data support a model in which *Ms5788* is regulated by translational
225 attenuation through *Ms5788A*-controlled Shine-Dalgarno availability, while *Ms5789* is
226 likely regulated by polarity, due to the absence of elongating ribosomes in *Ms5788*.
227
228 **Cysteine-dependent regulation of *Ms5789* and *Ms5790* by *Ms5788A* occurs in a**
229 **chromosomal context**
230
231 To assess whether our reporter-supported model is valid in the native locus context, we
232 performed quantitative mass spectrometry-based proteomics (LFQ) to determine

233 differences in protein expression for cells grown with or without cysteine
234 supplementation. Whole cell extracts were prepared from cultures of *M. smegmatis*
235 grown in minimal media +/- cysteine supplementation. Tryptic digests of whole cell
236 lysates were subjected to nanoUHPLC-MS/MS identified and quantitated using label-free
237 based peak integration (Cox and Mann 2008; Bosserman et al. 2017). As expected, the
238 abundance of most proteins is unchanged between the two conditions (+/- cysteine),
239 reflected in the linear correlation on the diagonal of the scatter plot (Fig 4A). Proteins
240 below the diagonal were more abundant in cells grown without cysteine supplementation.
241 The small (159 AA) and hydrophobic Ms5788 was not detected in these experiments.
242 However, *Ms5789* and *Ms5790* are predicted to be co-transcribed in an operon with
243 *Ms5788* (Martini et al. 2019), and expression of the encoded proteins Ms5789 and
244 Ms5790 was higher in cells grown without cysteine (Fig 4A, yellow diamonds),
245 consistent the attenuation model.
246
247 To test the hypothesis that polycysteines in *Ms5788A* control Ms5789 and Ms5790
248 expression, we generated an out-of-frame (OoF) mutation in *Ms5788A* by deletion of a
249 single nucleotide from the chromosomal locus. This deletion shifts the reading frame to
250 encode valines and alanines in place of the polycysteine tract, and adds six additional
251 amino acids (ERSRAL) prior to encountering a stop codon (Fig 4B), while preserving the
252 potential for nearly wild-type RNA duplex formation. This mutant exhibited low Ms5789
253 and Ms5790 expression levels consistent with a stable mRNA leader structure
254 sequestering the Shine-Dalgarno of *Ms5788*. Expression of Ms5789 and Ms5790 was

255 unaffected by cysteine supplementation (Fig 4B, yellow diamonds in diagonal),
256 highlighting the role of cysteine codons in relieving attenuation.
257
258 We reasoned that if *Ms5788A* directs attenuation, its deletion would elevate operon
259 expression of the downstream genes. We created a precise deletion of *Ms5788A* and
260 subjected this mutant to LFQ proteomics. As predicted, Ms5789 and Ms5790 were no
261 longer responsive to ambient cysteine supplementation, appearing in the diagonal of
262 unresponsive genes (Fig 4C, yellow diamonds in diagonal). Moreover, absolute levels of
263 Ms5789 and Ms5790 were higher in the Δ*5788A* strain than in either the wild-type or the
264 OoF mutant, regardless of cysteine supplementation. This indicates that the *Ms5788A*
265 deletion leaves the *Ms5788* Shine-Dalgarno sequence fully available for canonical
266 translation initiation, and results in an elevated expression of the operonic Ms5789 and
267 Ms5790 (Fig 4B, C yellow diamonds). Thus, data from wild-type and targeted
268 chromosomal mutant *M. smegmatis* are in agreement with our reporter-generated data,
269 and strengthen the conclusion that the *Ms5788A* LL-sORF modulates operonic gene
270 expression through an attenuation mechanism.
271
272 **Widespread cysteine-dependent regulation in *M. smegmatis* associated with cysteine-**
273 **rich LL-sORFs**
274
275 Given that LL-sORFs in *M. smegmatis* are enriched for polycysteine, we hypothesized
276 that additional cysteine-responsive genes are regulated by an associated polycysteine LL-
277 sORF. We identified six more LL-sORFs that contain at least two consecutive cysteine

278 codons and are located upstream of annotated genes. We then examined our proteomic
279 data for the protein levels encoded by putative operonic genes downstream of these LL-
280 sORFs for cells grown with/without cysteine supplementation. Remarkably, all of the
281 detectable proteins that were encoded by polycysteine LL-sORF-led operons were
282 upregulated, falling below the diagonal (Fig 4A, black circles), consistent with
283 attenuation release. In cysteine-replete medium, expression of some of these proteins
284 (Ms0113, Ms0934, Ms4527, Ms5279, and Ms5280) was below the threshold of detection
285 for reliable quantification [$<10^5$ LFQ intensity (a.u.)], indicative of tight attenuation. The
286 polycysteine LL-sORFs exhibit RNA-seq and Ribo-seq expression profiles consistent
287 with their robust expression during cysteine replete growth (Extended Data Figs 1 and 3).
288 None of these expressed LL-sORFs were identified by genome annotation pipelines.
289 Each of these cysteine-responsive operons contains genes annotated for cysteine
290 associated activities (Table 1). Collectively, these data reveal a cysteine-metabolic
291 regulon, whose concerted response is controlled independently at each locus by an
292 expressed LL-sORF that includes consecutive cysteine codons. This bottom-up
293 coordinated regulation contrasts with the conventional top-down, master regulator-driven
294 mechanism of transcriptional regulons.
295
296 LFQ proteomics does not comprehensively identify all proteins in a proteome. We looked
297 for additional loci with the same hallmarks of responsive attenuation: an expressed
298 polycysteine LL-sORF upstream of annotated operonic genes. *Ms4536A* was identified
299 upstream of a single gene (Table 1, Extended Data Fig 3 E). Without corroborating
300 indicators of cysteine response or function of the encoded operon protein, we only

301 speculate at its membership in the *M. smegmatis* cysteine LL-sORF regulon. The
302 independent evolution of LL-sORFs makes it unlikely that our experimental reference
303 species, *M. smegmatis*, contains all of the LL-sORF operons in the mycobacterial pan-
304 genome. In one example, transcriptomic profile data for *M. tuberculosis* clearly identified
305 *Rv2334A* as an unannotated gene meeting all of the criteria demonstrated in the *M.*
306 *smegmatis* regulon: an expressed LL-sORF with a C-terminal polycysteine tract,
307 followed by genes annotated as *cysK1* and *cysE* (Table 1, Extended Data Fig 3 G).
308
309 The availability of complete genome sequence information on diverse mycobacteria
310 provided an opportunity to track the evolution of these polycysteine sORFs. We searched
311 the genomes of 41 species using complementary approaches predicated on the sequence
312 of each *M. smegmatis* or *M. tuberculosis* LL-sORF, or on the position of the flanking
313 annotated orthologous genes as landmarks. We identified genomic sequences consistent
314 with conservation of the LL-sORFs expressed in *M. smegmatis* and *M. tuberculosis*
315 (Extended Data Table 2). The distribution of the conserved LL-sORFs is summarized as
316 barcodes adjacent to each species on the *Mycobacterium* genus phylogenetic tree (Fig 5).
317 Ms4527A, Ms4533A, and Ms5788A are deeply rooted, indicating both an emergence that
318 predates mycobacterial radiation and a selective advantage to retaining these sequences.
319 The presence/absence of others (e.g., *Ms0932A* or *Ms5280A*) is consistent with horizontal
320 gene transfer or sporadic gene loss by deletion or degradation. Sequence logos derived
321 from multiple alignments of the amino acid sequences encoded by the LL-sORFs further
322 support the evolutionary selection of their consecutive cysteine tracts (Fig 5). The
323 similarity between Ms4536A and Rv2334A strongly suggests homology by common

324 origin, yet the context, including operon genes, is not homologous. The occurrence of
325 Ms4536A—TQXA or Rv2334A—cysK1-cysE is mutually exclusive, suggesting origin
326 by rearrangement rather than a merodiploid duplication or horizontal gene transfer event.
327 Non-cysteine amino acids are also conserved in some LL-sORFs, suggesting that they
328 provide function through ribosomal interaction, or support activities as an independent
329 small protein product, or are encoded by codons that are constrained at the nucleotide
330 level.
331
332 DISCUSSION
333
334 **An attenuation mechanism that controls translation in response to amino acid**
335 **availability**
336 In the work presented here, we demonstrate attenuation of a cysteine biosynthesis pathway
337 locus. Attenuation is a recurring theme in biosynthetic pathways for nucleosides and amino
338 acids, in which the end product of the pathway interacts with the 5' leader of the
339 biosynthesis-encoding transcript to reduce (attenuate) expression of the operonic ORFs
340 downstream (Turnbough 2019). However, these models typically involve the formation of
341 competing alternate hairpin structures that function as an intrinsic terminator when the end
342 product is plentiful. By contrast, the *Ms5788A* LL-sORF featured here defines a class of
343 translational attenuator that indirectly assesses charged $tRNA^{cys}$ availability to modulate
344 expression of the downstream operonic genes.
345
346 **The need for cysteine regulation in mycobacteria**

347 Why might mycobacteria need a multi-locus cysteine regulon? Mycobacteria do not
348 produce glutathione, which modulates redox balance in most bacteria. Instead, they rely on
349 mycothiol (MSH), a cysteine derivative (Xu et al. 2011; Loi et al. 2015). The *mshA* gene
350 (*Ms0933*) encodes the first enzyme in MSH biosynthesis, and it resides in an operon with
351 the hallmarks of cysteine attenuation (Extended Data Fig 3 B). The multiple roles of
352 cysteine in mycobacterial protein synthesis, redox and sulfur metabolism likely require
353 subtle, independent fine-tuning of the enzymes involved in these respective pathways. The
354 slight differences in cysteine composition and architectures of the polycysteine LL-sORFs
355 could impart varying mechanisms as well as customized levels of operon gene expression.
356

357 **Attenuating sORF requirements**

358 What are the cysteine codon requirements of an effective small ORF attenuator? Our
359 criterion that LL-sORFs encode two consecutive cysteine codons may seem to be a low
360 threshold, yet the *cis*-encoded proteins only detectable in cysteine-limiting conditions
361 (baseline of Fig 4A) demonstrate the effectiveness of two consecutive cysteines in LL-
362 sORFs in relieving attenuation. Our Ms5788A LL-sORF analysis demonstrated that base
363 pairing was needed to impose attenuation, but it is premature to speculate that it is required
364 at all responsive loci. Transcription activation was not a regulatory factor at the *Ms5788*
365 locus; the promoter remained intact in the *Ms5788A* mutants in Fig 4 B, C, indicating that
366 all of the cysteine response at this locus is directed by the attenuation mechanism we
367 detailed and not via transcriptional activation. RNA-seq profiles of the LL-sORFs
368 presented here are also consistent with robust constitutive transcription in cysteine-replete
369 medium.

370

371 **Evolution of a cysteine attenuation regulon**

372 The independent evolution of similar polycysteine LL-sORF architecture associated with
373 cysteine attenuation at multiple loci in mycobacteria indicates that coordinated expression
374 is important, and that LL-sORF directed attenuation is effective. The co-regulation of
375 individual operons functionally defines a regulon, akin to regulons controlled by dedicated
376 DNA-binding transcription factors. We speculate that the evolution of regulation by
377 attenuation is simplified in mycobacteria by the robust nature of leaderless translation, and
378 that the evolution of a dedicated transcription factor and its cognate binding sites in the
379 promoters of target operons is more problematic than exploiting LL-sORFs in a genus that
380 exhibits frequent and robust LL-mRNA expression. As the first genes in their transcripts,
381 LL-mRNAs are ideally positioned to *cis*-regulate expression of downstream genes.
382 Additionally, transcription start sites are preferentially associated with purines (R), and the
383 +2 transcript position is preferentially associated with pyrimidines (Y) (Martini et al.
384 2019). Thus, transcription often begins at RYN trinucleotide sequences. Our previous study
385 showed that a 5' RUG trinucleotide is both necessary and sufficient for robust leaderless
386 translation initiation (Shell et al. 2015), so many transcription start sites are predicted to
387 already initiate leaderless translation, and others are only one or two changes from
388 initiating translation. It is not yet clear whether leaderless architecture *per se* offers
389 advantages for attenuation and may have been selected over canonical Shine-Dalgarno
390 translation initiation, or whether the sole criterion of an RUG sequence at the transcription
391 start site is simply a relatively modest requirement. It is clear, however, that in
392 mycobacteria, polycysteine LL-sORFs integrate two levels of coordination: locally by

393 modulating polycistronic operon gene expression, and globally by synchronizing operon
394 response.
395
396 **Concluding remarks**
397 Given the prevalence of LL-sORFs in mycobacteria, we speculate that other translational
398 regulons have evolved as an alternative to transcriptional regulons controlled by DNA-
399 binding transcription factors. LL translation is considered to be the ancestral form of
400 ribosome delivery (Nakamoto 2009; Zheng et al. 2011; Duval et al. 2013), indicating that
401 LL-sORF regulons may be ancient and widespread. LL-sORFs define a functional
402 subclass of small *de novo* genes that effectively decode translational stress into broad
403 regulatory effects.
404
405 METHODS
406
407 Bacterial strains and culture
408 *M. smegmatis* wild-type mc^2155 and its derivatives were grown in tryptic soy broth +
409 0.05% Tween 80 (TSBT) or on TSA plates, and cultured at 37˚C. Antibiotic selection for
410 reporter maintenance or mutation selection strategies included apramycin (12.5 µg/ml on
411 agar, 10 µg/ml in broth), hygromycin (100 µg/ml and 25 µg/ml), kanamycin (50 µg/ml
412 and 10 µg/ml), and zeocin (50 µg/ml and 25 µg/ml).
413
414 For the cysteine attenuation study, bacteria were cultured in minimal media. Base
415 medium per liter: 6g Na_2HPO4 (anhydrous), 3g KH_2PO4, 0.5g NaCl, 1g NH_4Cl, 0.05%

416 (v/v) Tween-80. After autoclaving, 0.2 % glucose and micronutrients were added to final
417 concentrations: $MgSO_4$ to 1 mM, $CaCl_2$ to 100 µM, H_3BO_3 to $4x10^{-7}$ M, $CoCl_2.6H_2O$ to
418 $3x10^{-8}$ M $CuSO_4.5H_2O$ to $1x10^{-8}$ M, $MnCl_2.4H_2O$ to $8x10^{-8}$ M, $ZnSO_4.7H_2O$ to $1x10^{-8}$
419 M, $FeSO_4.7H_2O$ to $1x10^{-6}$ M. L-cysteine or L-cystine (exogenously stable dimeric
420 cysteine) was supplemented at 200 µg/ml as noted.
421
422 Luciferase assays
423 Reporters were generated by long-primer-dimer PCR to recreate the LL-sORF and leader
424 sequence of *M. smegmatis msmeg_5788*. The products were cloned by Infusion and
425 verified by DNA sequence analysis. The NanoLuc (Promega) luciferase ORF is carried
426 on a plasmid that confers apramycin resistance and integrates at the L5 *attB* site in
427 mycobacteria. Mycobacterial cultures were grown in minimal media for luciferase assays.
428 Luciferase activity in a culture was assessed by the addition of NanoGlo (Promega)
429 substrate and then measuring luminescence normalized to culture density.
430
431 Chromosomal mutants of *M. smegmatis*
432 A precise deletion of *Ms5788A* was created using a targeting plasmid that integrates via a
433 single cross-over allowing selection of a hyg^r intermediate and, then, is resolved by a
434 second homology-driven event in which a *sacB*/*galK* counter-selection allows enrichment
435 for the deleted recombinant (Barkan et al. 2011). The OoF point mutant was created by a
436 recombineering approach that used a single-stranded oligonucleotide template to
437 introduce an additional adenine in *Ms5788A* (van Kessel and Hatfull 2008). Co-
438 electroporation of 2×10^{-10} mol of 60-mer oligo with 500 ng of an episomal zeo^r plasmid

439 (pGE324, Zeo-sacB) allowed zeocin selection and isolation of the electrocompetent
440 population of *M. smegmatis*. Mismatch-sensitized PCR (MAMA-PCR) screening (Cha et
441 al. 1992; Swaminathan et al. 2001) of isolates identified clones that integrated the
442 additional adenine. Deletion and point mutants were verified by genomic DNA PCR and
443 sequencing.
444
445 Mass spectrometry
446 Wild-type and mutant derivative *M. smegmatis* were cultured in minimal media with or
447 without cysteine supplementation, harvested by centrifugation and cryo-milled (Retsch
448 MM400, Haan, Germany) for mass spectrometry analysis. Reagents were of LC-MS
449 quality or higher and obtained from Sigma Aldrich unless indicated. Milled cell pellets
450 were digested with trypsin using commercial S-Traps (Protifi, NY). Briefly 50 μg of
451 protein from each milled cell pellet was re-suspended in 6% SDS 10mM Tris-2-Carboxy
452 ethyl phosphine in 100 mM tri ethyl ammonium bicarbonate (TEAB), heated at 95 ºC for
453 3minutes, then alkylated with 10 mM iodoacetamide in the dark for 15 min. Samples
454 were acidified by addition of H_3PO4_4 to 1.2% Final (v/v) and flocculated by 7-fold
455 addition of 95/5 MeOH:100 mM TEAB prior to collection on the S-trap column
456 (Zougman et al. 2014). Washing and conditioning was preformed three times by 150 μl
457 addition of MeOH buffer as above and 1 μg of sequencing grade trypsin (Promega, WI)
458 was added to each sample and digested at 37 ºC for 8 hours. Peptides were isolated,
459 acidified and desalted using Stage tips packed into P100 pipette tips, and dried using a
460 MiVac (Genvac UK) prior to LC-MS/MS analysis (Rappsilber et al. 2003).
461

462 NanoUHPLC-MS/MS was performed essentially as described (Bosserman et al. 2017;
463 Bosserman et al. 2019). 1μg of each digest was analyzed in technical triplicate and
464 biological duplicate on an Orbitrap instrument running a TOP15 data-dependent
465 acquisition (Q-Exactive Thermo San Jose, CA). Protein spectral matching and Label Free
466 quantification were performed using MaxQuant (Cox and Mann 2008) against the *M.*
467 *smegmatis* FASTA combined with contaminants from the Uniprot database, LFQ param
468 (UP000000757 6,595 entries). LFQ parameters were set to default, quantification was
469 restricted to proteins >2 peptides (missing peaks enabled). Target-decoy was used to
470 determine False Discovery Rates (Elias and Gygi 2007) and proteins at a global 1% FDR
471 were used for quantification. Data reduction and significance testing were performed
472 using a modified LIMMA methodology (Efstathiou et al. 2017). Protein search and RAW
473 data files are accessible at the Center for Computational Mass Spectrometry via
474 ftp://MSV000084381@massive.ucsd.edu (Deutsch et al. 2017). Proteins quantitatively
475 replicated in one condition but absent in another, were given an arbitrary intensity of
476 $5x10^4$ (below detection threshold) for ease in visualization.
477
478 Bioinformatics
479
480 Phylogeny
481 The genomes and proteomes for *Mycobacterium abscessus* subsp. massiliense
482 (NC_018150.2), *M. africanum* strain 25 (CP010334,1), *M. litorale* strain F4
483 (CP019882.1), *M. avium* 104 (NC008595.1), *M. ulcerans* Agy99 (NC_008611.1), *M.*
484 *vanbaalenii* PYR-1 (NC_008726.1), *M. marinum* M (NC_010612.1), *M. liflanddii*

485 128FXT (NC_020133.1), *M. kansasii* ATCC 12478 (NC_022663.1), *M. gilvum* Spyr1
486 (NC_014814.1), *M. bovis* AF2122/97 (NC_002945.4), *M. tuberculosis* H37Rv
487 (NC_000963.3), *M. sinense* strain JDM601 (NC_015576.1), *M. canettii* CIPT 140010059
488 (NC_015848.1), *M. chubuense* NBB4 (NC_018027.1), *M. intracellulare* MOTT-64
489 (NC_016948.1), *M. intracellulare* ATCC 13950 (NC_016946.1), *M. neoaurum* VKM ac-
490 1815D (NC_023036.2), *M. haemophilum* ATCC 29548 (NZ_CP0118883.2), *M. simiae*
491 ATCC 25275 (NZ_HG315953.1), *M. goodii* strain X7B (NZ_CP012150.1), *M. fortuitum*
492 strain CT6 (NZ_CP011269.1), *M. phlei* strain CCUG 21000 (NZ_CP014475.1), *M.*
493 *immunogenum* strain CCUG 47286 (NZ_CP011530.1), *M. chelonae* CCUG 47445
494 (NZ_CP007220.1), *M. vaccae* 95051 (NZ_CP011491.1), *M. chimaera* strain AH16
495 (NZ_CP012885.2), *M. caprae* strain Allgaeu (NZ_CP016401.1), *M. colombiense* CECT
496 3035 (NZ_CP020821.1), *M. dioxanotrophicus* strain PH-06 (NZ_CP020809.1), *M.*
497 *marseillense* strain FLAC0026 (NZ_CP023147.1), *M. lepraemurium* strain Hawaii
498 (NZ_CP021238.1), *M. shigaense* strain UN-152 (NZ_AP018164.1), *M. stephanolepidis*
499 (NZ_AP018165.1), *M. pseudoshottsii* JCM 15466 (NZ_AP018410.1), *M. paragordonae*
500 49061 (NZ_CP025546.1), *M. rutilum* strain DSM 45405 (NZ_LT629971.1), *M.*
501 *thermoresistibile* strain NCTC10409 (NZ_LT906483.1), *M. hassiacum* DSM 44199
502 (NZ_LR026975.1), and *M. microti* strain 12 (CP010333.1) were downloaded from NCBI.
503 Orthologs of *M. smegmatis* proteins were identified by a reciprocal best blast hit
504 approach and aligned in MAFFT v.7.058b (Katoh and Standley 2013). A supermatrix
505 was created from the concatenation of 1560 of these protein alignments with an in-house
506 Python script. Alignments included in the supermatrix were required to have 40-41 of
507 the 41 mycobacterial species present. IQ-TREE v.1.6.9 (Nguyen et al., 2015) generated a

508 maximum likelihood (ML) phylogeny under an LG+R4 model of evolution (Soubrier et
509 al., 2012; Yang 1995; Le and Gascuel 2008). Support values were generated by the
510 ultrafast bootstrap method with 1000 replicates (Minh et al., 2013). The ML tree was
511 visualized in FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).
512
513 Identification of ssORFs in additional mycobacterial genomes
514 The orthologs to all *M. smegmatis* ORFs with an upstream ssORF were identified in other
515 Mycobacteria via a reciprocal best blast hit approach and the regions 1000 nucleotides
516 upstream of these genes were extracted. *M. smegmatis* ssORF proteins were queried
517 against these upstream regions via TBLASTN searches to identify orthologous
518 sequences. The coordinates for these sequences were obtained from the blast results and
519 extended to the match the length of the corresponding *M. smegmatis* ssORF. To ensure
520 that we captured the start and stop codon, we extended these coordinates by an additional
521 5-10 amino acids on both 5' and 3' ends. These coordinates and the strand information
522 captured from the blast results were used to extract and translate ssORFs from 1 Kb
523 upstream regions with an in-house Python script. All ssORFs were visually inspected
524 and initially aligned in MEGA v. 70.26 (Kumar et al., 2016). In some cases, additional
525 ssORF orthologs were identified by predicting proteins (anything between two stop
526 codons) in upstream 1 Kb regions with the getorf subroutine from EMBOSS v. 6.3.1
527 (Rice et al., 2000).
528

529 Multiple alignments of deduced LL-sORF amino acid sequences were performed using a
530 web-based tool and default clustal algorithm (Madeira et al. 2019). The aligned
531 sequences were compiled into a sequence logo infographic (weblogo.berkeley.edu).
532

533 Supplementary Information Table 1. LL-sORFs predicted by transcriptomic data: Ribo-

534 seq, RNA-seq, and RUG transcription start sites (Shell et al. 2015; Martini et al. 2019).

535 REFERENCES

536

537 Barkan D, Stallings CL, Glickman MS. 2011. An improved counterselectable
538 marker system for mycobacterial recombination using galK and 2-deoxy-
539 galactose. *Gene* **470**: 31-36.
540 Bechhofer DH. 1990. Triple post-transcriptional control. *Mol Microbiol* **4**: 1419-
541 1423.
542 Beck HJ, Moll I. 2018. Leaderless mRNAs in the Spotlight: Ancient but Not
543 Outdated! *Microbiol Spectr* **6**.
544 Bosserman RE, Nguyen TT, Sanchez KG, Chirakos AE, Ferrell MJ, Thompson CR,
545 Champion MM, Abramovitch RB, Champion PA. 2017. WhiB6 regulation of
546 ESX-1 gene expression is controlled by a negative feedback loop in
547 Mycobacterium marinum. *Proc Natl Acad Sci U S A* **114**: E10772-E10781.
548 Bosserman RE, Nicholson KR, Champion MM, Champion PA. 2019. A New ESX-1
549 Substrate in Mycobacterium marinum That Is Required for Hemolysis but
550 Not Host Cell Lysis. *J Bacteriol* **201**.
551 Cha RS, Zarbl H, Keohavong P, Thilly WG. 1992. Mismatch amplification
552 mutation assay (MAMA): application to the c-H-ras gene. *PCR Methods*
553 *Appl* **2**: 14-20.
554 Cortes T, Schubert OT, Rose G, Arnvig KB, Comas I, Aebersold R, Young DB.
555 2013. Genome-wide mapping of transcriptional start sites defines an
556 extensive leaderless transcriptome in Mycobacterium tuberculosis. *Cell*
557 *Rep* **5**: 1121-1131.
558 Couso JP, Patraquim P. 2017. Classification and function of small open reading
559 frames. *Nat Rev Mol Cell Biol* **18**: 575-589.
560 Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates,
561 individualized p.p.b.-range mass accuracies and proteome-wide protein
562 quantification. *Nat Biotechnol* **26**: 1367-1372.
563 Crappe J, Van Criekinge W, Trooskens G, Hayakawa E, Luyten W, Baggerman G,
564 Menschaert G. 2013. Combining in silico prediction and ribosome
565 profiling in a genome-wide search for novel putatively coding sORFs.
566 *BMC Genomics* **14**: 648.
567 Deana A, Belasco JG. 2005. Lost in translation: the influence of ribosomes on
568 bacterial mRNA decay. *Genes Dev* **19**: 2526-2533.
569 Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T,
570 Campbell DS, Bernal-Llinares M, Okuda S, Kawano S et al. 2017. The
571 ProteomeXchange consortium in 2017: supporting the cultural change in
572 proteomics public data deposition. *Nucleic Acids Res* **45**: D1100-D1106.
573 Duval M, Korepanov A, Fuchsbauer O, Fechter P, Haller A, Fabbretti A, Choulier
574 L, Micura R, Klaholz BP, Romby P et al. 2013. Escherichia coli ribosomal
575 protein S1 unfolds structured mRNAs onto the ribosome for active
576 translation initiation. *PLoS biology* **11**: e1001731.
577 Efstathiou G, Antonakis AN, Pavlopoulos GA, Theodosiou T, Divanach P,
578 Trudgian DC, Thomas B, Papanikolaou N, Aivaliotis M, Acuto O et al.

579 2017. ProteoSign: an end-user online differential proteomics statistical
580 analysis platform. *Nucleic Acids Res* **45**: W300-W306.
581 Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence
582 in large-scale protein identifications by mass spectrometry. *Nat Methods*
583 **4**: 207-214.
584 Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P,
585 Hayashizaki Y, Bailey TL, Grimmond SM. 2006. The abundance of short
586 proteins in the mammalian proteome. *PLoS Genet* **2**: e52.
587 Hemm MR, Paul BJ, Miranda-Rios J, Zhang A, Soltanzad N, Storz G. 2010. Small
588 stress response proteins in Escherichia coli: proteins missed by classical
589 proteomic studies. *J Bacteriol* **192**: 46-58.
590 Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. 2008. Small membrane
591 proteins found by comparative genomics and ribosome binding site
592 models. *Mol Microbiol* **70**: 1487-1501.
593 Henkin TM, Yanofsky C. 2002. Regulation by transcription attenuation in
594 bacteria: how RNA provides instructions for transcription
595 termination/antitermination decisions. *Bioessays* **24**: 700-707.
596 Hinnebusch AG, Ivanov IP, Sonenberg N. 2016. Translational control by 5'-
597 untranslated regions of eukaryotic mRNAs. *Science* **352**: 1413-1416.
598 Hobbs EC, Fontaine F, Yin X, Storz G. 2011. An expanding universe of small
599 proteins. *Curr Opin Microbiol* **14**: 167-173.
600 Loi VV, Rossius M, Antelmann H. 2015. Redox regulation by reversible protein S-
601 thiolation in bacteria. *Front Microbiol* **6**: 187.
602 Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey
603 ARN, Potter SC, Finn RD et al. 2019. The EMBL-EBI search and sequence
604 analysis tools APIs in 2019. *Nucleic Acids Res* **47**: W636-W641.
605 Martini MC, Zhou Y, Sun H, Shell SS. 2019. Defining the Transcriptional and
606 Post-transcriptional Landscapes of Mycobacterium smegmatis in Aerobic
607 Growth and Hypoxia. *Front Microbiol* **10**: 591.
608 Meydan S, Marks J, Klepacki D, Sharma V, Baranov PV, Firth AE, Margus T, Kefi
609 A, Vazquez-Laslop N, Mankin AS. 2019. Retapamulin-Assisted Ribosome
610 Profiling Reveals the Alternative Bacterial Proteome. *Mol Cell* **74**: 481-
611 493 e486.
612 Miranda-CasoLuengo AA, Staunton PM, Dinan AM, Lohan AJ, Loftus BJ. 2016.
613 Functional characterization of the Mycobacterium abscessus genome
614 coupled with condition specific transcriptomics reveals conserved
615 molecular strategies for host adaptation and persistence. *BMC Genomics*
616 **17**: 553.
617 Nakamoto T. 2009. Evolution and the universality of the mechanism of
618 initiation of protein synthesis. *Gene* **432**: 1-6.
619 Oppenheim DS, Yanofsky C. 1980. Translational coupling during expression of
620 the tryptophan operon of Escherichia coli. *Genetics* **95**: 785-795.
621 Rappsilber J, Ishihama Y, Mann M. 2003. Stop and go extraction tips for matrix-
622 assisted laser desorption/ionization, nanoelectrospray, and LC/MS
623 sample pretreatment in proteomics. *Anal Chem* **75**: 663-670.

624 Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, Pavlopoulos
625 GA, Kyrpides NC, Bhatt AS. 2019. Large-Scale Analyses of Human
626 Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* **178**: 1245-
627 1259 e1214.
628 Shell SS, Wang J, Lapierre P, Mir M, Chase MR, Pyle MM, Gawande R, Ahmad R,
629 Sarracino DA, Ioerger TR et al. 2015. Leaderless Transcripts and Small
630 Proteins Are Common Features of the Mycobacterial Translational
631 Landscape. *PLoS Genet* **11**: e1005641.
632 Swaminathan S, Ellis HM, Waters LS, Yu D, Lee EC, Court DL, Sharan SK. 2001.
633 Rapid engineering of bacterial artificial chromosomes using
634 oligonucleotides. *Genesis* **29**: 14-21.
635 Turnbough CL, Jr. 2019. Regulation of Bacterial Gene Expression by
636 Transcription Attenuation. *Microbiol Mol Biol Rev* **83**.
637 van Kessel JC, Hatfull GF. 2008. Efficient point mutagenesis in mycobacteria
638 using single-stranded DNA recombineering: characterization of
639 antimycobacterial drug targets. *Mol Microbiol* **67**: 1094-1107.
640 Weaver J, Mohammad F, Buskirk AR, Storz G. 2019. Identifying Small Proteins
641 by Ribosome Profiling with Stalled Initiation Complexes. *mBio* **10**.
642 Xu X, Vilcheze C, Av-Gay Y, Gomez-Velasco A, Jacobs WR, Jr. 2011. Precise null
643 deletion mutations of the mycothiol synthesis genes reveal their role in
644 isoniazid and ethionamide resistance in Mycobacterium smegmatis.
645 *Antimicrob Agents Chemother* **55**: 3133-3139.
646 Yanofsky C. 1981. Attenuation in the control of expression of bacterial operons.
647 *Nature* **289**: 751-758.
648 Zheng X, Hu GQ, She ZS, Zhu H. 2011. Leaderless genes in bacteria: clue to the
649 evolution of translation initiation mechanisms in prokaryotes. *BMC*
650 *Genomics* **12**: 361.
651 Zougman A, Selby PJ, Banks RE. 2014. Suspension trapping (STrap) sample
652 preparation method for bottom-up proteomics analysis. *Proteomics* **14**:
653 1006-1000.
654

655

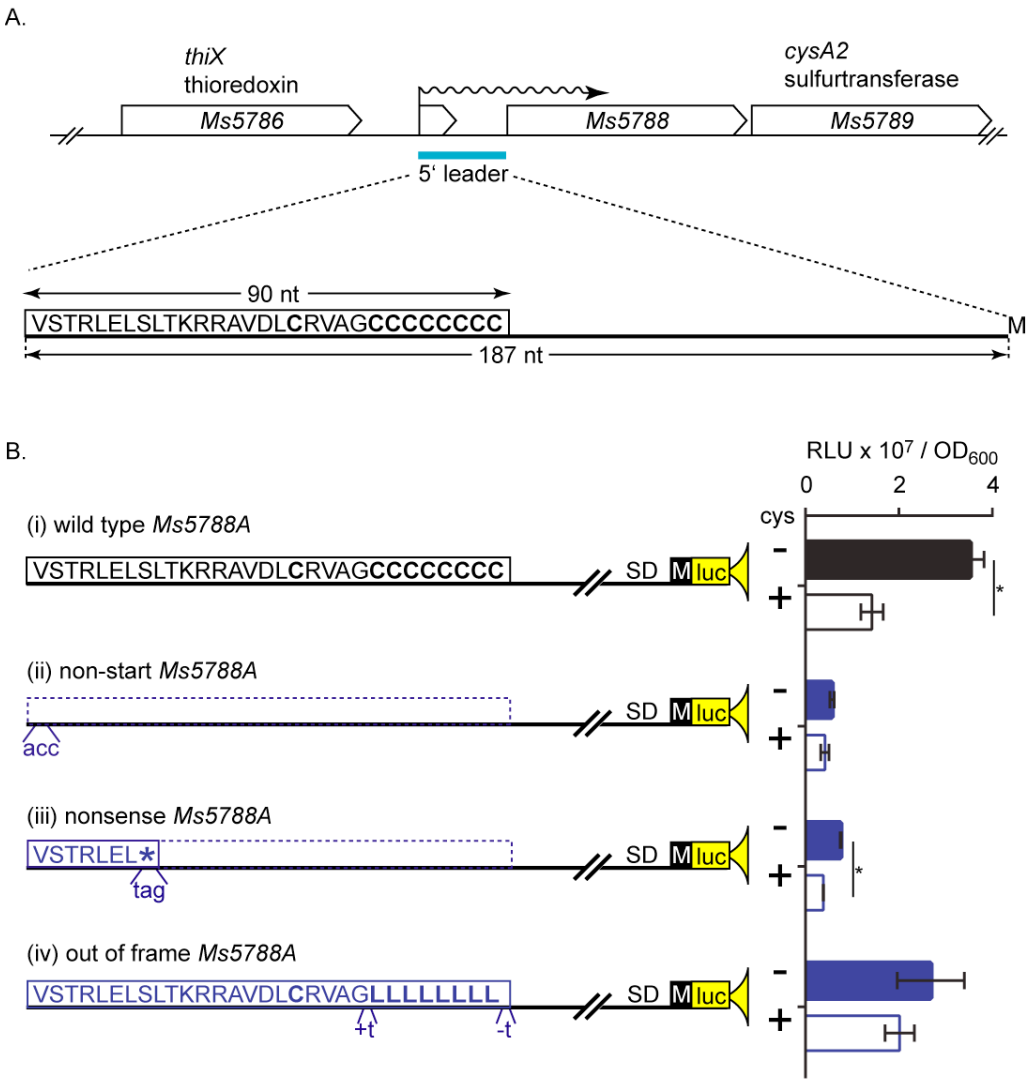


Figure 1. Polycysteine-encoding LL-sORF attenuates expression of operonic genes. **A.** Locus schematic of *Ms5788A* upstream of *Ms5788-5790*. The conceptual amino acid sequence of the wild-type Ms5788A shows 9 cysteine codons (bold C), with eight consecutive codons at the 3' end. **B.** Ribosome occupancy of *Ms5788A* relieves attenuation of luciferase reporter gene expression in the absence (-) of cysteine. The same wild-type (i) reporter data are shown in Figs 2-3 as a comparative reference for the mutant derivatives. Non-start (ii), nonsense (iii), and frameshift (iv) mutations decrease basal or attenuated expression (or both) of the luciferase gene. Lower case letters below

665 clones indicated nucleotide substitutions and insertion/deletions (+/-) used to create

666 mutants. Asterisks indicate significance $p < .01$ for cysteine supplementation, by two-

667 tailed t-test in Figs 1-3.

668
669

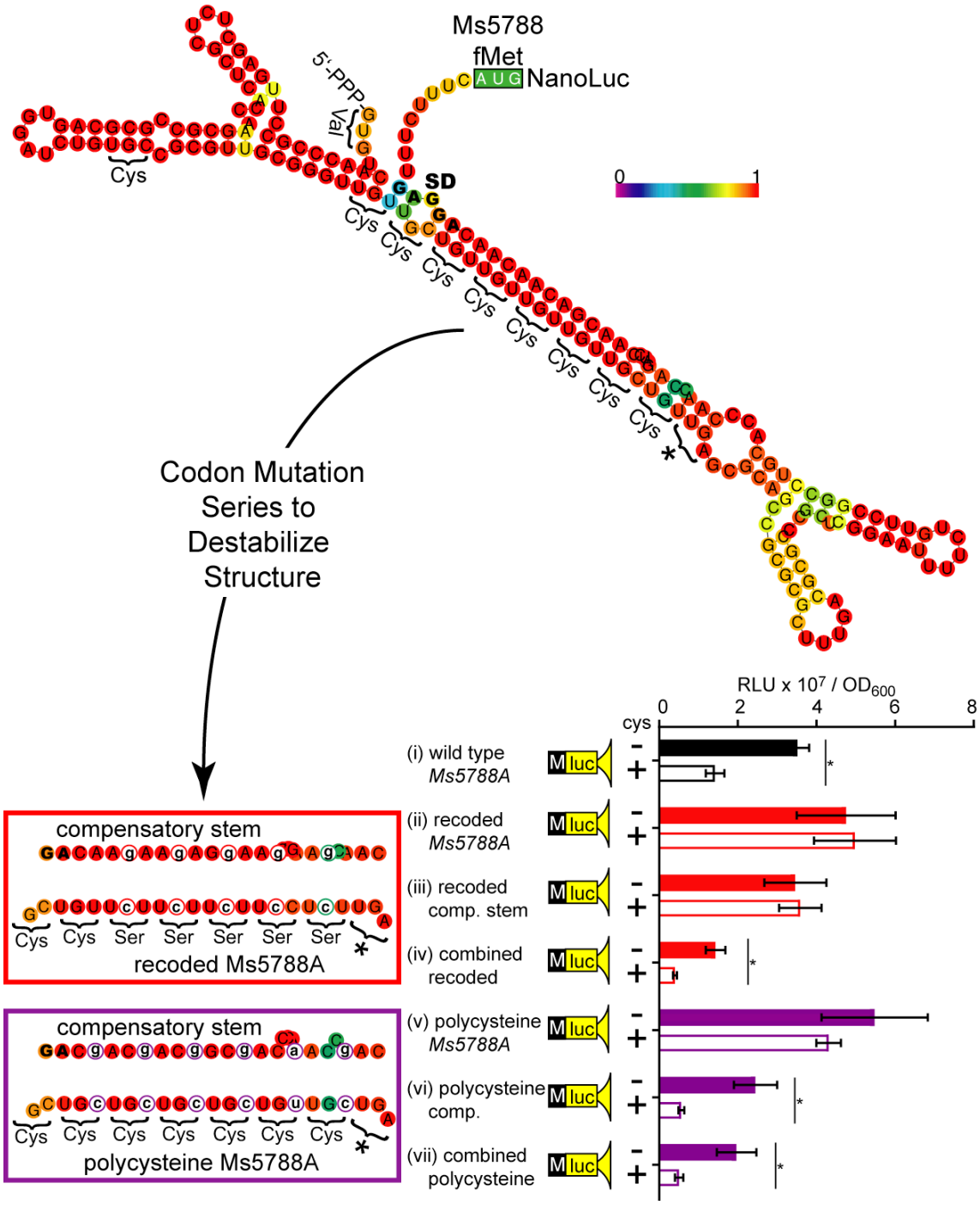


Figure 2. Predicted duplex mRNA structures regulate Shine-Dalgarno availability for expression of *Ms5788* luciferase reporter gene. A stable mRNA structure model (RNAWebSuite/RNAfold) shows extensive base pairing forming a stem between the polycysteine coding region of the LL-sORF and the peri-Shine-Dalgarno region. The

676 color of each nucleotide indicates the confidence in the predicted structure (as indicated
677 in the heatmap gradient key). Clustered point mutations were introduced into the
678 Ms5788A LL-sORF or the predicted pairing nucleotides to disrupt base pairing but, when
679 combined, should restore the modeled stable mRNA structure when combined. Two
680 series of mutants were generated (red, ii – iv and purple, v – vii). The effect of structure-
681 destabilizing mutations and re-stabilizing combination were tested for their effect on the
682 luciferase reporter fused to *Ms5788.*
683

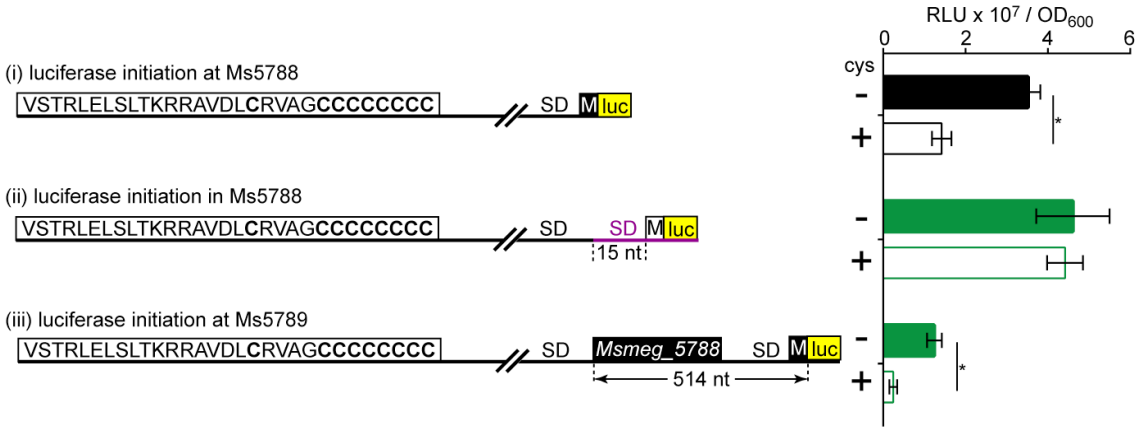


Figure 3. Attenuation regulates translation initiation of Ms5788. Nluc translated via an independent Shine-Dalgarno and initiation codon (M in white box) gages mRNA extension beyond the structured leader (ii). Placement of the luciferase reporter at the initiation codon of Ms5789 (iii) indicates that translation attenuation in Ms5788 has polar effects, propagating the effects of *Ms5788A*.

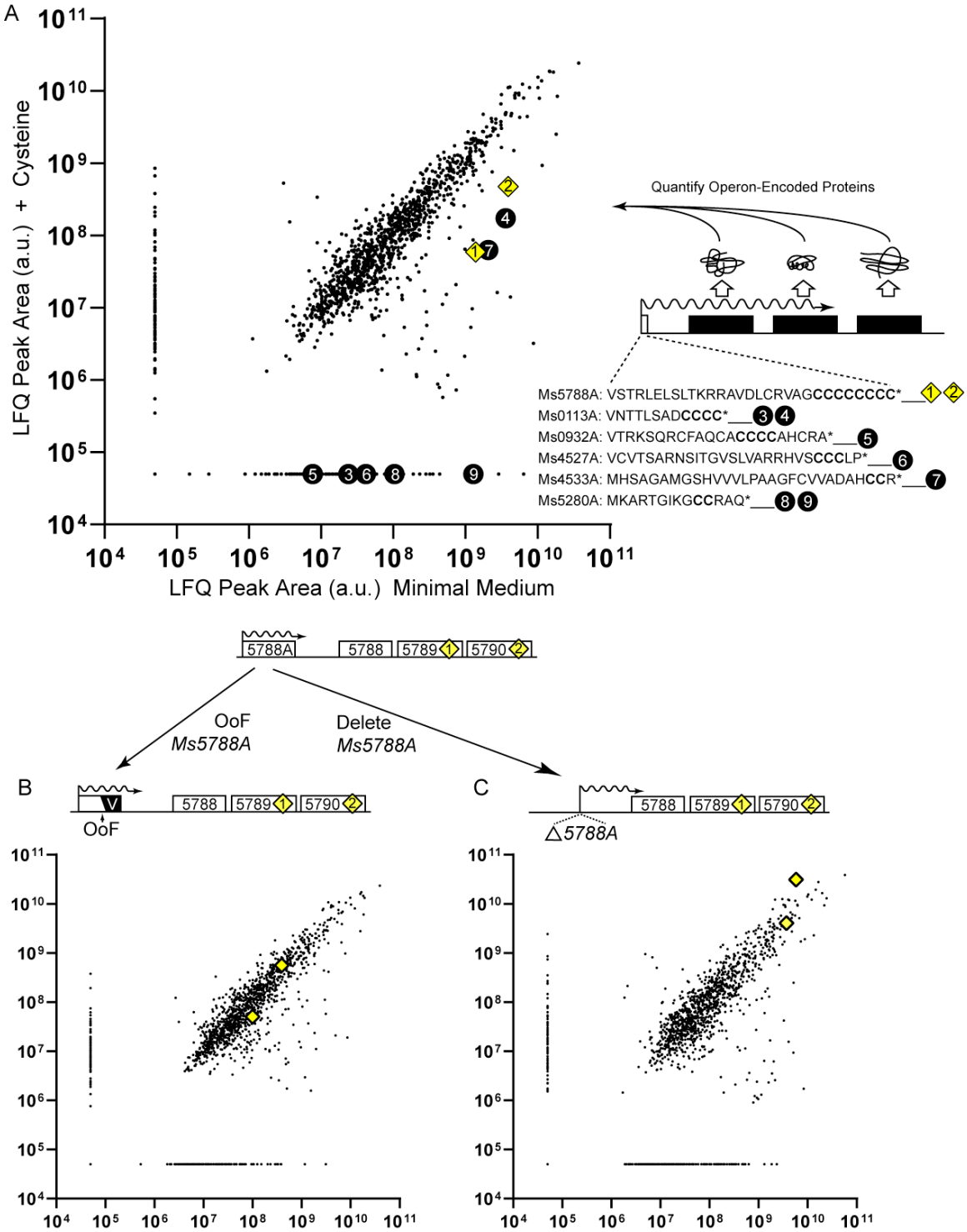


Figure 4. Cysteine attenuation is observed in the native context. a. Label-Free Quantitative proteomics (LFQ) was used to identify and quantitate changes in the abundance of proteins from whole-cell extracts of wild-type *M. smegmatis* subjected to

696 trypsin digestion and nanoUHPLC-MS/MS. *M. smegmatis* was cultured in minimal
697 media (X-axis) or supplemented with cysteine (Y-axis). Units are normalized LFQ peak
698 area (protein) in arbitrary units. The baselines for both X and Y-axes were artificially set
699 at $5x10^4$ counts to allow depiction of proteins not expressed in one of the two conditions.
700 Peptides from Ms5789 and Ms5790 (yellow diamond 1 and 2, respectively) increased in
701 abundance (shift right along the X-axis) under cysteine limitation. Peptides of annotated
702 proteins from loci similarly encoded on polycysteine-encoding LL-sORF mRNAs are
703 indicated by black filled circles (Fig S3; 3 = Ms0113; 4 = Ms0114; 5 = Ms0934; 6 =
704 Ms4527; 7 = Ms4533; 8 = Ms5279; 9 = Ms5280). b. A single nucleotide deletion caused
705 a frameshift mutation (OoF) in *Ms5788A* and changes the CCCCCCCC* to
706 VAVVVVAVersral*. This OoF LL-sORF mutant is insensitive to cysteine and does not
707 release Ms5789 and Ms5790 from attenuation. c. Deleting *Ms5788A* elevates expression
708 of Ms5789 and Ms5790 and is insensitive to cysteine, indicating a completely
709 unattenuated state.
710
711

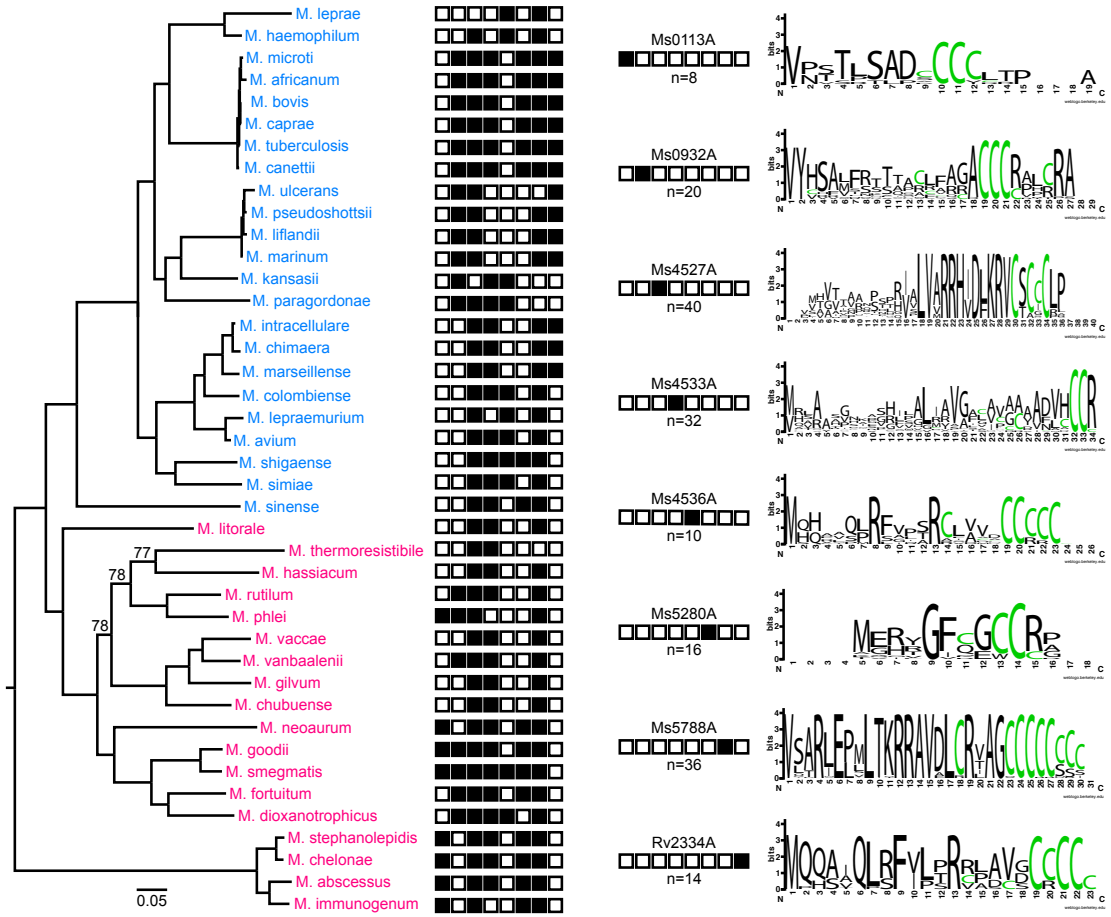
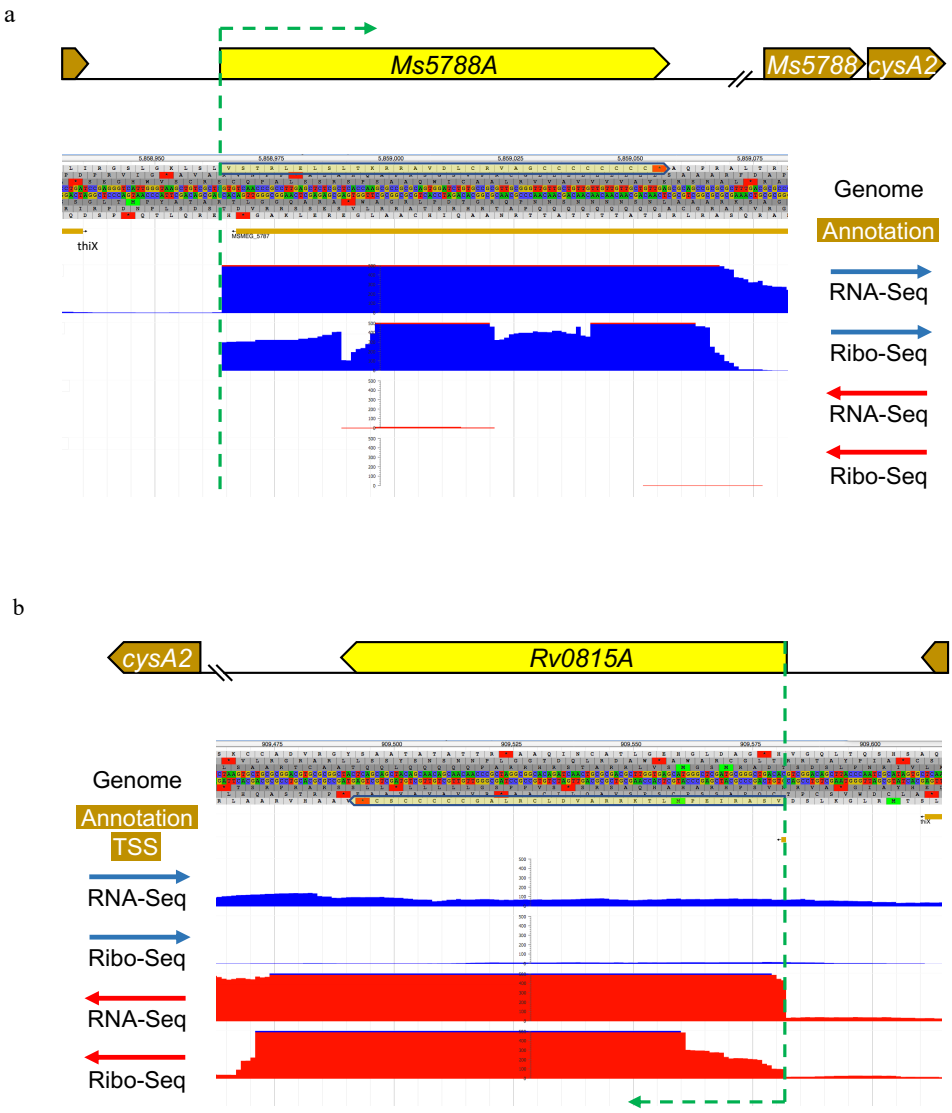


Figure 5. Phylogenetic distribution of polycysteine LL-sORFs in mycobacteria. A robust reference phylogenetic tree was constructed from complete genome sequences of 41 diverse *Mycobacterium* spp. All nodes had bootstrap values of 100, except where indicated. Species of the slow grower clade appear in blue text, and fast growers in magenta. The presence (black square) or absence (white square) of each LL-sORF forms a binary barcode for each species. Barcode key and fixed-length sequence logo for each LL-sORF is shown.

gene #	gene name	encoded protein predicted function	HMM Pfam	uniprot
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
Ms0932		ROK family protein	PF00480	A0QQZ7
Ms0933	mshA	MshA glycosyltransferase	PF00534	A0QQZ8
Ms0934		YbjN domain-containing protein		A0QQZ9
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
Ms4531	cysW	sulfate ABC transporter, permease protein	PF00528	A0R0W5
Ms4532	cysT	sulfate ABC transporter, permease protein CysT	PF00528	A0R0W6
Ms4533		sulfate ABC transporter substrate-binding protein	TIGR00971	A0R0W7
Ms4536		TQXA domain-containing protein; surface-expressed		A0R0X0
Ms5280		cysteine dioxygenase type I superfamily protein	PF05995	A0R2Y8
Ms5279		sulfurtransferase		A0R2Y7
Ms5788		DUF4395 domain-containing protein		A0R4C8
Ms5789	cysA2	putative thiosulfate sulfurtransferase	PF00581	A0R4C9
Ms5790		DUF1416 domain-containing protein, SseC protein	PF07210	A0R4D0
Rv2334	cysK1	O-acetylserine sulfhydrylase; Interacts with CysE		P9WP55
Rv2335	cysE	serine acetyltransferase; acetylates serine to make cysteine		P95231

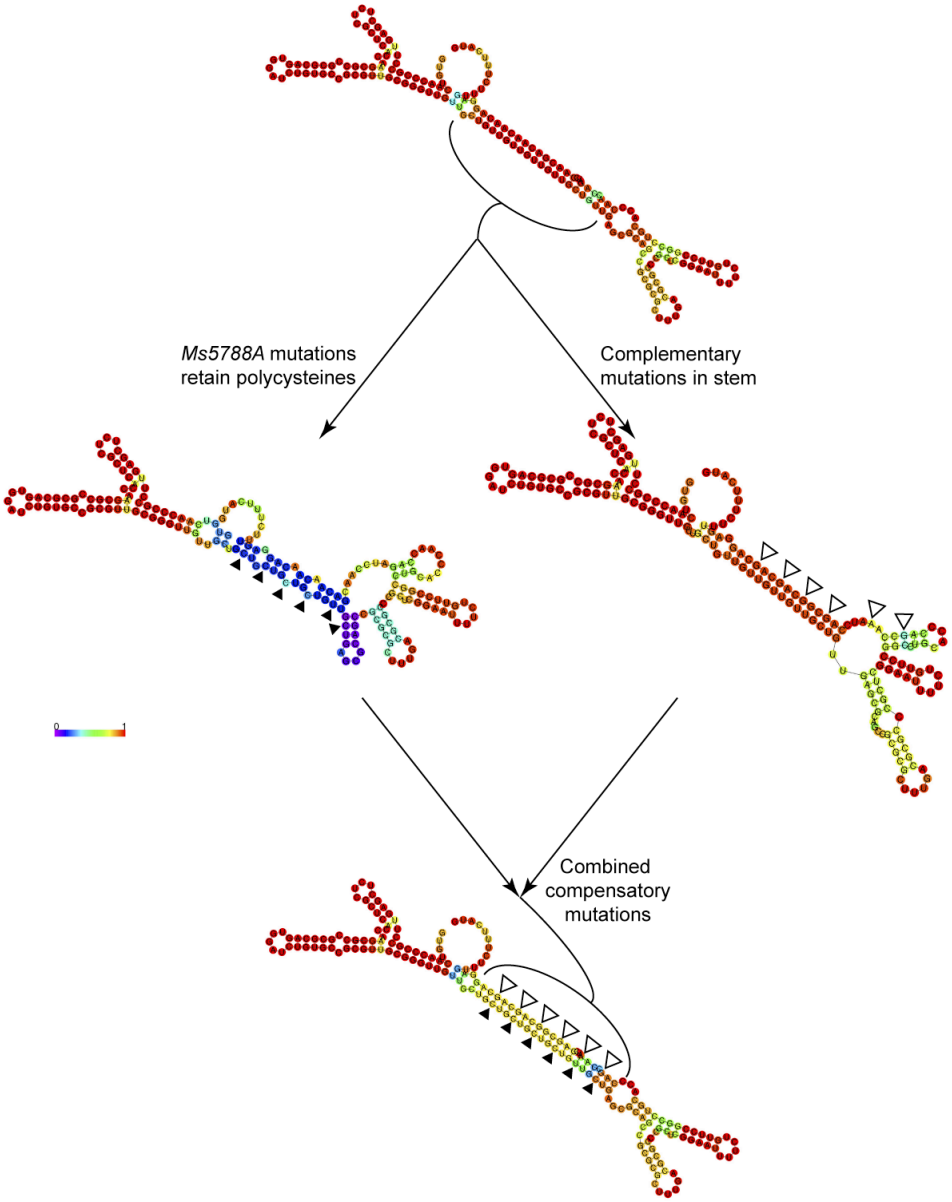
722
723
724 Table 1. Annotated genes of the cysteine translation attenuation regulon. Gene identifiers
725 are listed for *M. smegmatis* (Ms) or *M. tuberculosis* (Rv). The predicted functions and
726 features of the encoded proteins are shown, as annotated in Pfam or uniport record.
727
728

729 EXTENDED DATA FILES
730



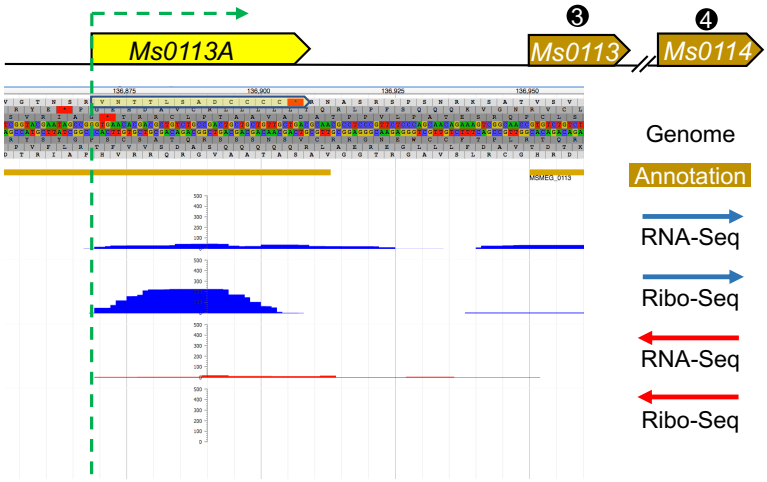
Extended Data Fig 1. a. Schematic and expression profile of the *Ms5788A* locus. RNA-seq and Ribo-seq profiles show robust occupancy of the positive strand (maximum displayed read depth = 500), corresponding to the LL-sORF (shaded yellow). The end of the transcriptionally independent *thiX* gene is shown, and *cysA2* (Ms5789) represents operon genes downstream. Reads from a single replicate are shown for space efficiency. Note that *Ms5787* (gold line shown as an annotated gene reading from right to left) is likely a misannotation, as we detect no transcriptional or translational activity for this anti-sense gene. b. Schematic and expression profile for the *M. tuberculosis* orthologous locus, *Rv0815A*.

731

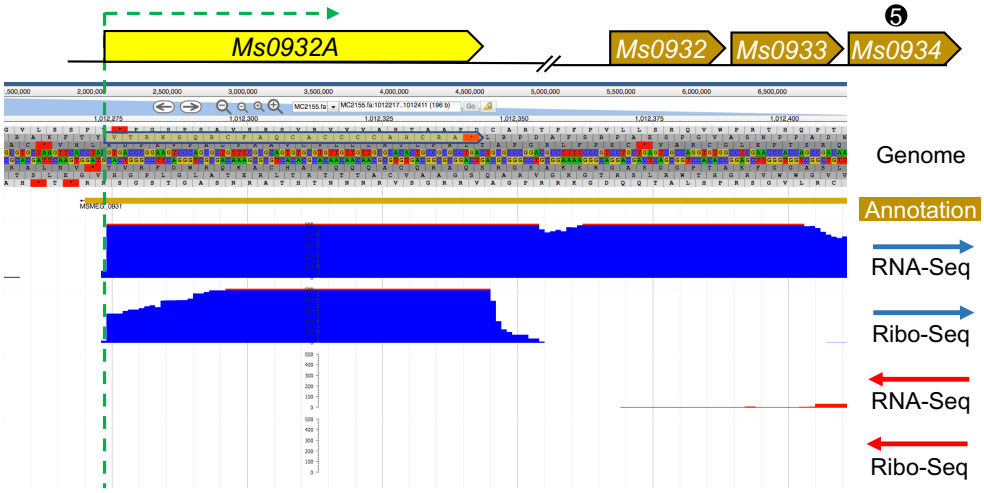


Extended Data Fig 2. RNA-folding model of the polycysteine *Ms5788A* mutant. The *Ms5788A* ORF mutations (indicated by filled triangles) retain polycysteine coding, but disrupt the stability and folding of the wild-type mRNA as indicated in the modeled structures. Mutations of the complementary nucleotides (indicated by open triangles) near the *Ms5788* Shine-Dalgarno sequence, are predicted to form G-U wobble base pairs; the resulting overall structure and stability are comparable to that of the wild-type mRNA. The comparable wild-type predicted folding of the complementary stem mutant is consistent with the comparable wild-type expression of the reporter (Fig 2 vi). Combining each mutated stem restores the predicted mRNA 2° structure and cysteine attenuation (Fig 2 vii). Color scheme is as for Fig 2.

3a



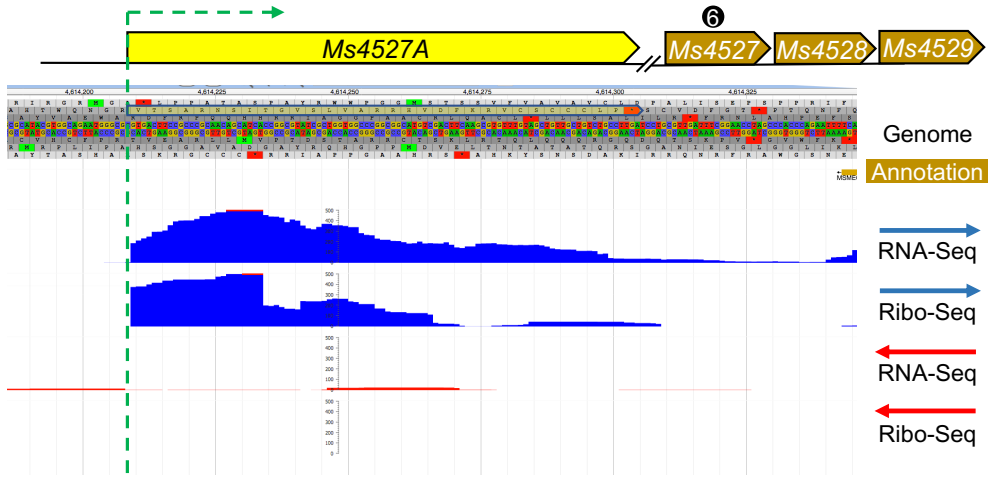
3b



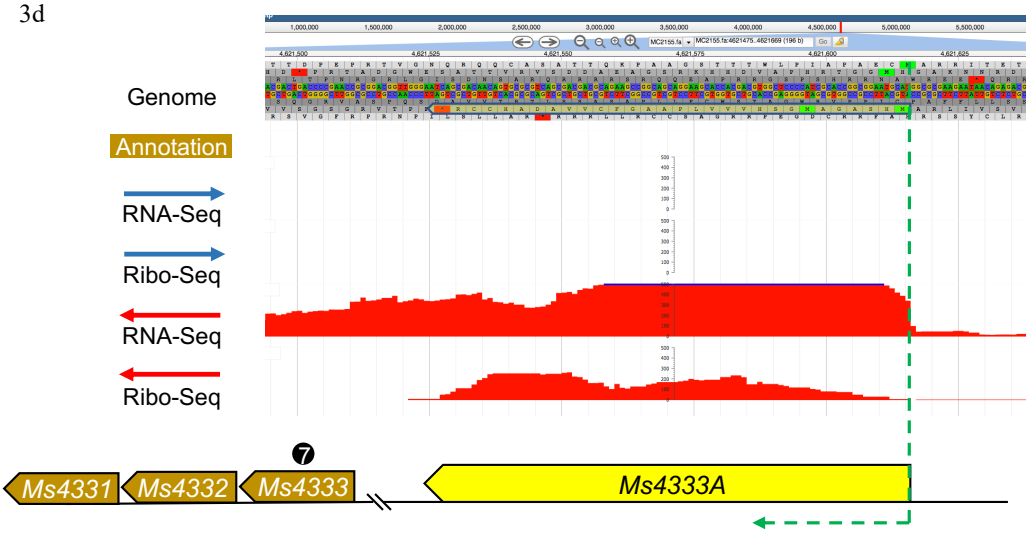
Extended Data Fig 3. Additional loci that feature expressed polycysteine LL-sORFs upstream of annotated genes. JBrowse images of each LL-sORF locus show the gene context and sequence and transcriptomic profiles. Yellow arrow boxes indicate the LL-sORF in each locus, and the gene encoding the cysteine-responsive protein detected by mass spectrometry is indicated by the black circle with white number as in the mass spectrometry scatter plot (Fig 5a). Panels are shown for (a) *Ms0113A/0013/Ms0114*, (b) *Ms0932A/0932/0933/0934.* Note that genes (B) *Ms0931* and (C) *Ms4528* are likely misannotated based on our transcriptional and translational data. Pipeline prediction algorithms annotated none of these LL-sORFs. Screen shots captured from http://www.wadsworth.org/research/scientific-resources/interactive-genomics.

734

3c



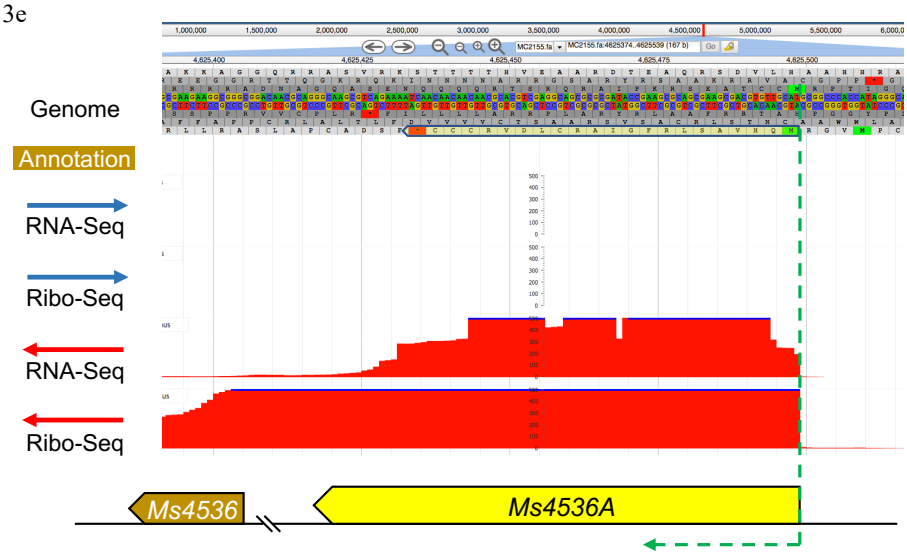
3d



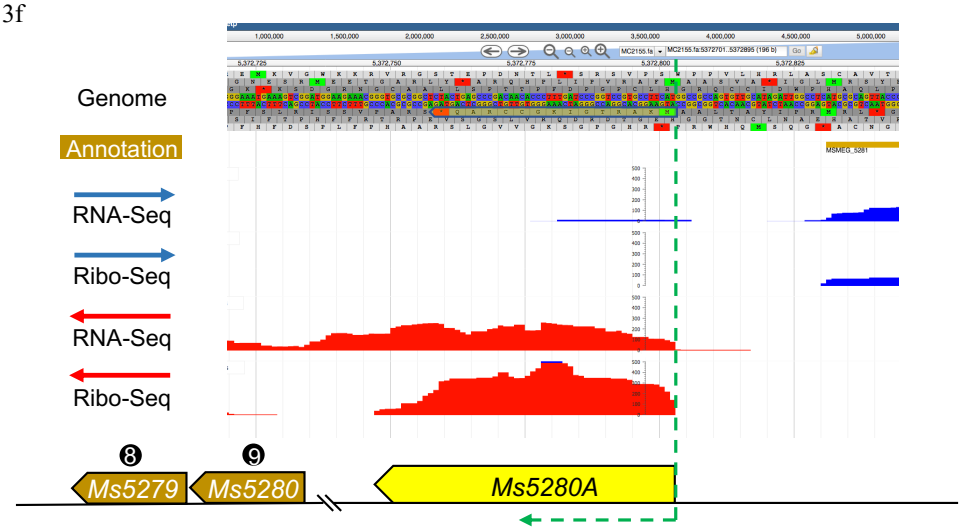
(C) *Ms4527A/4527/4528/4529*, (D) *Ms4533A/4533/4532/4531/4530*

735

3e



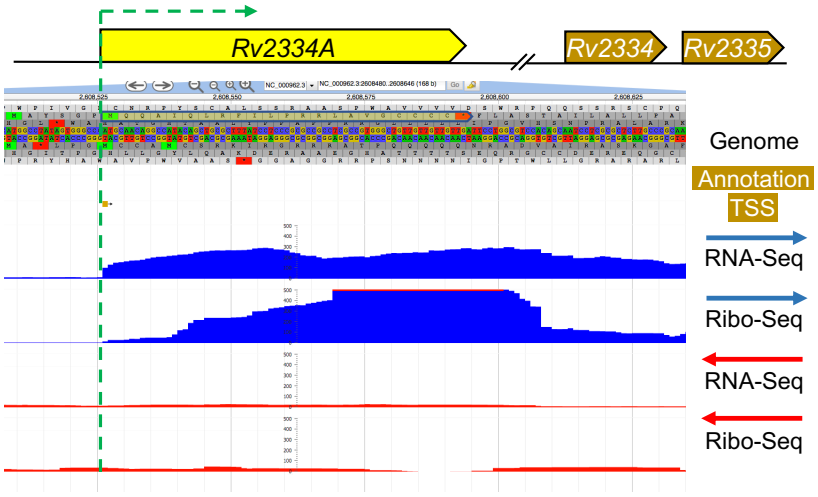
3f



(e) *Ms4536A/4536*, (f) *Ms5280/5279*

736

3g



(g) *Rv2334A*

737

738

ANNOTATED																								
Amino Acid	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**x10**	**x11**	**x12**	**x13**	**x14**	**x15**	**x16**	**x17**	**x18**	**x19**	**freq**	**exp x2**	**obs >1x en**	**>x1 inst**	**total AA**
A, alanine	197695	29314	4605	729	186	33	1	1	1											0.118932	27659.338	1.2606954	34870	232565
C, cysteine	16898	197	1	1																0.008743	149.48337	1.3312517	199	17097
D, asparate	115764	7902	627	26	4		6													0.063581	7904.9284	1.0835013	8565	124329
E, glutatmate	103393	5090	167	6																0.055566	6037.5465	0.8717117	5263	108656
F, phenylalanine	61161	2052	73	1		1														0.032365	2048.3106	1.0384167	2127	63288
G, glycine	155845	13140	1222	123	25	5	2			1			1							0.087123	14842.557	0.9782007	14519	170364
H, histidine	45231	1405	54	2	1															0.023878	1114.9531	1.3112659	1462	46693
I, isoleucine	85428	3059	82	1																0.045294	4011.6806	0.7832129	3142	88570
K, lysine	43100	1377	37	3																0.022766	1013.4559	1.3981862	1417	44517
L, leucine	168666	16936	1363	158	9	1	1													0.095699	17908.469	1.0312439	18468	187134
M, methioine	41595	881	16	3																0.021732	923.48263	0.9745717	900	42495
N, asparagine	44477	1187	33																	0.023369	1067.8947	1.1424347	1220	45697
P, proline	107885	5629	544	103	33	19	11	2	3	1				1						0.058417	6672.9984	0.9509968	6346	114231
Q, glutamine	56801	2368	115	5	3		1													0.030322	1797.8767	1.3860795	2492	59293
R, arginine	127562	11444	1220	138	15	6	3		1											0.071794	10079.041	1.2726409	12827	140389
S, serine	99161	5508	365	38	8	1	3		1											0.05374	5647.2176	1.0490122	5924	105085
T, threonine	114519	7025	569	42	10	4	4	1	2		1								1	0.062481	7633.7703	1.003305	7659	122178
V, valine	151768	14766	1590	156	10	1														0.086062	14483.544	1.1408119	16523	168291
W, tryptophan	29995	632	10																	0.015667	480.00475	1.3374868	642	30637
Y, tyrosine	42884	1038	19	1																0.022472	987.44451	1.0714526	1058	43942
TOTALS	1809828	130950	12712	1536	304	71	32	4	8	2	1	0	1	1	0	0	0	0	1	1	118632.23	21.84775	145623	1955451
LL-sORFs																								
Amino Acid	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**x10**	**x11**	**x12**	**x13**	**x14**	**x15**	**x16**	**x17**	**x18**	**x19**	**freq**	**exp x2**	**obs >x1 en**	**>x1 inst**	**total AA**
A, alanine	437	40	3																	0.09548	38.070312	1.129489	43	526
C, cysteine	142	14	2	2				1												0.034852	5.0724457	3.7457277	19	192
D, asparate	204	15																		0.042476	7.5343651	1.9908778	15	234
E, glutatmate	212	11	1																	0.043021	7.7287923	1.5526358	12	237
F, phenylalanine	103	4	1																	0.020693	1.7882352	2.7960527	5	114
G, glycine	407	31																		0.085133	30.26639	1.0242384	31	469
H, histidine	159	2																		0.029588	3.655865	0.5470661	2	163
I, isoleucine	158	3																		0.029769	3.7008599	0.8106224	3	164
K, lysine	88	2	1																	0.017245	1.24183	2.4157896	3	95
L, leucine	285	27	3																	0.063169	16.663777	1.8003122	30	348
M, methioine	90	4																		0.017789	1.3214998	3.0268639	4	98
N, asparagine	101																			0.018334	1.4036463	0	0	101
P, proline	331	17	3																	0.067889	19.246783	1.0391347	20	374
Q, glutamine	149	6																		0.029225	3.5667009	1.6822268	6	161
R, arginine	571	95	12	4																0.147577	90.948604	1.2204695	111	813
S, serine	419	35	5	3																0.093665	36.636531	1.1736919	43	516
T, threonine	318	25	2																	0.067889	19.246783	1.4028318	27	374
V, valine	283	27	4																	0.063351	16.759683	1.849677	31	349
W, tryptophan	114	2																		0.021419	1.915927	1.0438811	2	118
Y, tyrosine	63																			0.011436	0.54613	0	0	63
TOTALS	4634	360	37	9	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	176.92714	19.825789	407	5509

Extended Data Table 1. Amino acid content and clustering in proteins encoded by annotated and LL-sORF genes. Single or consecutive instances are totaled. Freq = specific AA/total AA, exp x2 = $(\text{Freq})^2$*total AA, obs >1x en = obs/exp, >1x inst = total of polyAA occurrences.

LL-sORF	0113A	0932A	4527A	4533A	4536A	5280A	5788A	Rv2334A
Mycobacterium_abscessus	Cys x 4		Cys x 1X2	Cys x 2		Cys x 2	Cys x 6	
Mycobacterium_africanum		Cys x 3	Cys x 1X1X1	Cys x 2		Cys x 2	Cys x 5X1	Cys x 4
Mycobacterium_avium_104			Cys x 1X1X1	Cys x 2	Cys x 6		Cys x 7	
Mycobacterium_bovis		Cys x 3	Cys x 1X1X1	Cys x 2		Cys x 2	Cys x 5 x 1	Cys x 4
Mycobacterium_canettii		Cys x 3	Cys x 1X1X1	Cys x 2		Cys x 2		Cys x 4
Mycobacterium_caprae		Cys x 3	Cys x 1X1X1	Cys x 2		Cys x 2	Cys x 5X1	Cys x 4
Mycobacterium_chelonae	Cys x 3		Cys x 1X2	Cys x 2		Cys x 2	Cys x 6	
Mycobacterium_chimaera			Cys x 1X1X1	Cys x 2			Cys x 7	Cys x 1 x 2
Mycobacterium_chubuense			Cys x 1X3	Cys x 4			Cys x 5X1	
Mycobacterium_colombiense			Cys x 1X1X1	Cys x 2	Cys x 7		Cys x 7	
Mycobacterium_dioxanotrophicus		Cys x 3	Cys x 1X3	Cys x 2	Cys x 4		Cys x 7	
Mycobacterium_fortuitum		Cys x 4	Cys x 1X3	Cys x 2		Cys x 2	Cys x 7	
Mycobacterium_gilvum			Cys x 1X3	Cys x 3			Cys x 7	
Mycobacterium_goodii	Cys x 4	Cys x 4	Cys x 1X3	Cys x 2		Cys x 2	Cys x 7	
Mycobacterium_haemophilum			Cys x 1X3		Cys x 3 x 2		Cys x 5X1	
Mycobacterium_hassiacum			Cys x 1X3	Cys x 3			Cys x 6	
Mycobacterium_immunogenum	Cys x 3		Cys x 1X2			Cys x 2	Cys x 7	
Mycobacterium_intracellulare			Cys x 1X1X1	Cys x 2			Cys x 7	Cys x 1 x 2
Mycobacterium_kansasii		Cys x 3		Cys x 2	Cys x 5			
Mycobacterium_leprae					Cys x 1 x 2		Cys x 1 x 1 x 1	
Mycobacterium_lepraemurium			Cys x 1X3		Cys x 6		Cys x 5	
Mycobacterium_liflandii		Cys x 3	Cys x 1X3				Cys x 5X1	Cys x 5
Mycobacterium_litorale			Cys x 1X3	Cys x 2			Cys x 6	
Mycobacterium_marinum_M		Cys x 3	Cys x 1X3				Cys x 5 x 1	Cys x 5
Mycobacterium_marseillense			Cys x 1X1X1	Cys x 2			Cys x 7	Cys x 3
Mycobacterium_microti		Cys x 3	Cys x 1X1X1	Cys x 2		Cys x 2	Cys x 5X1	Cys x 4
Mycobacterium_neoaurum	Cys x 4		Cys x 1X3	Cys x 3		Cys x 2	Cys x 5	
Mycobacterium_paragordonae		Cys x 3	Cys x 1X3	Cys x 2	Cys x 5 x 1			
Mycobacterium_phlei	Cys x 3	Cys x 4	Cys x 1X3				Cys x 6	
Mycobacterium_pseudoshottsii		Cys x 3	Cys x 1X3				Cys x 5X1	Cys x 5
Mycobacterium_rutilum		Cys x 3	Cys x 1X3	Cys x 4			Cys x 6	
Mycobacterium_shigaense			Cys x 1X1X1	Cys x 2	Cys x 2 x 2	Cys x 2	Cys x 2X3	
Mycobacterium_simiae			Cys x 1X1X1	Cys x 2	Cys x 2 x 2		Cys x 7 x 1	
Mycobacterium_sinense			Cys x 1X1X1	Cys x 2		Cys x 3	Cys x 4X1	
Mycobacterium_smegmatis	Cys x 4	Cys x 4	Cys x 1X3	Cys x 2	Cys x 3	Cys x 2	Cys x 8	
Mycobacterium_stephanolepidis	Cys x 3		Cys x 1X2	Cys x 2		Cys x 2	Cys x 6	
Mycobacterium_thermoresistibile		Cys x 3	Cys x 1X3					
Mycobacterium_tuberculosis		Cys x 3	Cys x 1X1X1	Cys x 2		Cys x 2	Cys x 5X1	Cys x 4
Mycobacterium_ulcerans		Cys x 3	Cys x 1X3					Cys x 5
Mycobacterium_vaccae			Cys x 1X3	Cys x 3			Cys x 6	
Mycobacterium_vanbaalenii		Cys x 4	Cys x 1X3	Cys x 3			Cys x 6	

Extended Data Table 2. Polycysteine LL-sORF gene conservation in mycobacteria. Each column designates the query LL-sORF. Open cells did not return an orthologous gene. Cys x describes the array of cysteines in the polycysteine cluster; arrays that are interrupted by non-cysteine codons represent that codon by X.