Title: Identification of new genes on a whole genome scale using saturated reporter transposon mutagenesis

Emily C. A. Goodall[1]*, Freya Hodges[1], Weine Kok[1], Budi Permana[1], Thom Cuddihy[1], Zihao Yang[1], Nicole Kahler[2], Kenneth Shires III[2], Karthik Pullela[1], Von Vergel L. Torres[1], Jessica L. Rooke[1], Antoine Delhaye[3], Jean-François Collet[3], Jack A. Bryant[4], Brian Forde[1], Matthew Hemm[2], Ian R. Henderson[1]

[1] Institute for Molecular Bioscience, University of Queensland, Brisbane, 4072, Australia

[2] Department of Biological Sciences, Towson University, Towson, USA

[3] Institut de Duve, UC Louvain, Brussels, Belgium

[4] Biodiscovery Institute, University of Nottingham, Nottingham, UK

*Corresponding author: e.goodall@imb.uq.edu.au

Key words: [up to 5, alphabetical order, separated by '/']

E. coli / gene expression / small proteins / transposon sequencing / translation reporter / tn-seq

Abstract

Small or overlapping genes are prevalent across all domains of life but are often overlooked for annotation and function because of challenges in their detection. The advent of high-density mutagenesis and data-mining studies suggest the existence of further coding potential within bacterial genomes. To overcome limitations in existing protein detection methods, we applied a genetics-based approach. We combined transposon insertion sequencing with a translation reporter to identify translated open reading frames throughout the genome at scale, independent of genome annotation. We applied our method to the well characterised species *Escherichia coli* and identified ~200 putative novel protein coding sequences (CDS). These are mostly short CDSs (<50 amino acids) and in some cases highly conserved. We validate the expression of selected CDSs demonstrating the utility of this approach. Despite the extensive study of *E. coli*, this method revealed proteins that have not been previously described, including proteins that are conserved and neighbouring functionally important genes, suggesting significant functional roles of small proteins that are still overlooked. We present this as a complementary method to whole cell proteomics and ribosome trapping for condition-dependent identification of protein CDSs. We anticipate this technique will be a starting point for future high-throughput genetics investigations to determine the existence of unannotated genes in multiple bacterial species.

Introduction

The last two decades have seen an substantial increase in sequenced bacterial genomes, coupled with an increase in the associated pan-genome of any given organism. However, in recent years there has been a growing body of work to suggest that our annotation of bacterial genomes is incomplete. First, smaller proteins have been historically overlooked in all domains of life [1–4]. Early definitions of a gene or coding sequence applied a size cut-off for annotation of a gene, ranging from 50-100 codons. However, several translated open reading frames (ORFs) as small

44 as two or three codons have been identified in *Escherichia coli* and other species [5–7]. A small
45 protein, as defined by Storz and colleagues, is directly translated from an ORF (as opposed to
46 processed from a larger protein), and composed of fewer than 50 amino acids [4]. Small proteins
47 (also referred to as "short ORF-encoded proteins" (SEPs), "microproteins" or "sproteins") have
48 been identified across the bacterial kingdom (reviewed extensively elsewhere [8–10]), with roles
49 in stress response (e.g. Prli42, TisB [11–13]), cell division (e.g. MciZ, SidA, Blr [14–17]), metabolism
50 (e.g. SgrT, MgtS, MntS [18,19]), respiration (e.g. CcoM, CydX [20,21]), signal transduction (e.g. Sda,
51 MgrB [22,23]), modulating antibiotic resistance (e.g. Blr, AcrZ, Prli53 [6,24]), defence and
52 toxin/antitoxin systems (PepA1, TisB, Fst [12]), or mediating host-interaction (e.g. rio3 [25]) and
53 survival inside host cells (KdpF [26]). Given their functional importance, a number of studies
54 have sought to identify these previously overlooked small proteins at scale [27], but this has not
55 been without its challenges. Small proteins are harder to predict bioinformatically as the short
56 sequence length makes it difficult to distinguish true coding sequences from random in-frame
57 sequences between a start and stop codons [28]. Furthermore, they are difficult to predict through
58 sequence conservation analysis as some small proteins have species-specific function and
59 therefore are genus or lineage-restricted.

60 Another class of overlooked genes are nested genes. Nested genes are defined as those
61 encoded within another gene. They can be encoded within the same orientation as the primary
62 coding sequence or anti-sense to the primary transcript. A subset of nested genes are isoforms[29],
63 which are encoded from an alternate start codon within the primary transcript, such as *cheA*(L)
64 and *cheA*(S) in *E. coli* [30], or the three isoforms of translation initiation factor 2 encoded by *infB*
65 [31,32]. Nested genes are prevalent in viruses and eukaryotes due to their compact genomes or
66 coding complexity, respectively, but considered rare anomalies in bacterial genomes [33]. As
67 such, many automated bacterial genome annotation tools discard putative coding sequences
68 that are within larger genes. However, well documented examples of (non-isoform) nested

genes include the *Shigella* enterotoxin (ShET) 1 gene encoded antisense to the Pic autotransporter [34], both virulence factors of diarrhoeagenic *E. coli* isolates, and *comS* encoded in the same sense but out of frame (OOF) to *srfA* in *Bacillus subtilis,* both of which have complementary but independent functions in natural competency and surfactin synthesis and secretion, respectively [35]. More recent discoveries of overlapping genes (*olg1* and *olg2* respectively) as large as 957 and 1728 nucleotides have been demonstrated in *Pseudomonas aeruginosa* [36], suggesting overlapping protein coding is not limited to just small proteins.

Due to the limitations outlined above, high-throughput methods to identify protein coding sequences are needed to fully elucidate the genetics and biology of a given organism. Specialist methods that trap mRNA-bound ribosomes (Ribo-seq) have been developed to identify translated mRNA and therefore protein coding sequences [29,37,38]. Although these studies highlight the remarkable extent to which small proteins have been overlooked, the data can however pose challenges for interpretation [39]. Ribo-seq methods have been further optimised through the use of antibiotics that trap translating ribosomes at specific positions along translated mRNA[40]; one example termed 'Ribo-Ret' uses retapamulin, which traps initiating ribosomes resulting in enrichment of ribosomes localised near the initiating start codon. Such adaptations improve the resolution of the output data, but pose other challenges such as differentiating true coding sequence from 'pervasive translation' [41]. Small proteins are also challenging to detect via mass spectrometry (discussed in detail by Ahrens *et al.* [42]). They might be transiently expressed, expressed only under specific conditions [43], low in abundance, have few sites for trypsin digestion, or highly hydrophobic membrane-associated proteins. Taken together, it is clear that complementary methods are needed if we are to fully realise the coding potential of any given organism.

Translation reporters are one tool for identification of protein coding sequences. Such assays involve the fusion of the C-terminus of the target protein to the N-terminus of an enzyme

94 or protein with a measurable read out, such as *β*-galactosidase, alkaline phosphatase or a
95 fluorescent protein. Most translation reporter assays have been applied to defined target
96 sequences, often in isolation. However, an early transposon-based random mutagenesis
97 approach introduced the *phoA* gene, encoding alkaline phosphatase, into a Tn*5* transposon
98 derivative 'Tn*phoA*'. The transposon was designed such that the alkaline phosphatase, which
99 is only functional in the periplasm, lacked its signal peptide sequence and therefore would only
100 be functional if fused to sequences that promote its export [44], thereby identifying secreted
101 proteins and periplasmically located portions of integral membrane proteins. This approach
102 pre-dated the high-throughput amplicon-sequencing technology available today but
103 demonstrates the utility of a transposon-based reporter system. Here, we combined a mini-Tn*5*
104 translation reporter with ultra-dense transposon mutagenesis and sequencing to query the *E.*
105 *coli* K-12 genome for new protein coding sequences independently of existing genome
106 annotation at scale. By using this approach, we identified previously uncharacterised protein
107 coding genes. We also document their conservation among >200,000 *E. coli* genomes. The
108 capacity to discern an organism's coding potential in a single experiment promises to be a
109 powerful strategy for unveiling genotype-phenotype relationships.
110

111 **Results**

112 ***Construction of a reporter-transposon library***

113 In our previous research[45], the resolution afforded from ultra-dense transposon mutagenesis
114 highlighted the presence of newly annotated small genes that contribute to cell fitness. We
115 hypothesised that there might be additional as yet unannotated coding sequences in the *E. coli*
116 genome, and that a transposon mutagenesis approach could be leveraged for their detection. A
117 reporter transposon was constructed by introducing a kanamycin resistance gene upstream of
118 the chloramphenicol resistance cassette of the transposon. However, the transposon was

119 designed such that the aminoglycoside-3'-phospotransferase (*aph*) gene that confers kanamycin
120 resistance was introduced immediately after the mini-Tn5 inverted repeat without a promoter,
121 ribosome binding site (RBS) or start codon (Fig. 1). Expression of the *aph* gene is dependent
122 upon in-frame insertion of the transposon into an actively translated gene. We first confirmed
123 that the intact antibiotic selection cassettes did not confer any cross resistance (S. Fig. 1). We
124 next verified that the *aph* gene could confer kanamycin resistance with an N-terminal fusion
125 of the transposon inverted repeat at the 5' end of the *aph* gene. The transposon was cloned in-
126 frame into LacZα in a pUC19 vector such that expression of *aph* was dependent upon Plac and
127 the start codon of LacZα. The LacZα-Tn fusion was sufficient to confer kanamycin resistance
128 and this resistance was lost if the LacZα start codon is mutated from ATG to AGG (S. Fig. 2).
129 Having confirmed the functionality of the transposon, we introduced the transposon via
130 transposition into the *E. coli* K-12 strain BW25113 genome at random aiming for 1 transposon
131 insertion event per cell. Transposon mutants were selected on LB agar plates supplemented
132 with chloramphenicol and were pooled to form the library. The transposon-genomic DNA
133 (gDNA) junctions were sequenced to identify the transposon insertion site. We designed our
134 sequencing primers to sequence from the kanamycin resistance gene into the gDNA upstream
135 to directly identify CDS translation fusion sites (S. Fig. 3). We sequenced two aliquots of the
136 library and obtained 4.2 M and 2.7 M sequencing reads per replicate respectively. Comparison
137 of the insertion density per gene between technical replicates showed sequencing the
138 transposon junction was reproducible between replicates with a Pearson correlation coefficient
139 of 0.99 (S. Fig. 4A), therefore we pooled the replicate data resulting in a total of ~6.9 M mapped
140 reads. Sub-sampling of our transposon junction sequences confirmed we had obtained
141 sufficient sequencing for full coverage of our library (S. Fig. 4B). We identified 185,709 and
142 184,405 unique insertions for each orientation of the transposon respectively representing a
143 total of 370,114 translational reporter mutants. When we further delineated the data by reading

144 frame there was an equivalent number of mutants for each frame (S. Fig. 4C), distributed
145 evenly throughout the genome (S. Fig.4D).
146
147 ***Identification of putative new genes***
148 To identify expressed protein coding sequences, we repeated selection of the library on agar
149 plates supplemented with an inhibitory concentration of kanamycin, with the expectation that
150 kanamycin resistance will only be conferred if the transposon has inserted in frame into an
151 expressed protein coding sequence. For selection of kanamycin resistant mutants, 25 µg/ml of
152 kanamycin was identified as the minimum concentration of kanamycin that inhibits growth of
153 *E. coli* BW25113 on supplemented LB agar plates while selecting for transposon mutants (S.
154 Fig. 5A-B). We chose to screen the library at this concentration to minimise the selection stress
155 and maximise the number of translation-fusion mutants recovered. The library was screened
156 twice, on separate days. We sequenced the transposon-gDNA junctions as before and recovered
157 >150,000 reads per sample, which mapped to ~8,000 unique insertion sites per replicate. The
158 number of recovered unique insertions was lower than the number of fusion mutants we
159 expected to recover. This may be a result of disrupted mRNA stability when endogenous CDSs
160 are disrupted mid-sequence: indeed, review of the relative position of insertion sites found they
161 were predominantly within the extreme 5' or 3' of each CDS (S. Fig. 6B). The biological
162 replicates of mutants selected on kanamycin plates had a strong Pearson correlation coefficient
163 (r = 0.81) between identified insertion sites (S. Fig. 6A). To account for the possibility that
164 individual mutants might grow if they acquire spontaneous kanamycin resistance, we filtered
165 the data to include hits that were identified in both independent screens, resulting in a final
166 number of 6,219 transposon insertion sites. We then reviewed the functionality of the screen:
167 we split the insertion data by reading frame and confirmed the bp level of resolution for
168 identification of translated reading frames. For example, within the *tig-clpP-clpX-lon-hupB*

169 operon the transposon insertions were each consistent with the appropriate reading frame for
170 translation fusion with the kan^{R} *aph* gene (Fig. 2), demonstrating the functionality of the screen.
171 Moreover, this level of resolution can reveal putative CDSs that do not coincide with annotated
172 genes, for example three unannotated putative novel CDSs within the *tig* operon (Fig. 2).

173 Of the conservative 6,219 insertion sites, 5,754 insertions (92.52%) were in-frame
174 within 1,139 annotated genes, 272 insertions (4.37%) were within annotated genes but out of
175 frame and 193 insertions (3.10%) were intergenic (between annotated protein coding
176 sequences; Fig. 3A). We used the well curated GenBank sequence (accession CP009273) as a
177 reference for protein coding genes, however, to account for more recently annotated genes we
178 also cross referenced our data against the RefSeq annotation (annotated by the NCBI
179 prokaryotic genome annotation pipeline [PGAP], accession NZ_CP009273, annotated 24-
180 APR-2023). This led to the positive identification of an additional 7 genes, 5 of which (*mgtT*,
181 *pssL*, *ynfU*, *yqgH* and *ysgD*) were identified in a targeted Ribo-Ret screen in *E. coli* K-12 strain
182 MG1655, further validating the utility of our approach to identify small coding sequences [38].
183 Of the remaining two, *rseD* and *nadS*, *rseD* has been previously validated by chromosomal
184 integration of a *lacZ* translation reporter [46]. While *nadS* (RS26030), although not currently
185 listed in Ecocyc, is annotated in the PGAP reference genome as a hypothetical protein, our data
186 presents the first lab-based evidence for its expression and suggests it is expressed under
187 normal laboratory growth conditions. Overall, the revised total of known genes that were
188 identified in the screen was 1,146. We cross-referenced our data against other whole-cell
189 protein-identification datasets, including a proteomic mass-spectrometry dataset and a Ribo-
190 Ret dataset [29,47], chosen for using a comparable strain (*E. coli* BW25113 or derivative) and
191 growth condition (LB +/- supplements; S. Table 1). The largest cohort of proteins were shared
192 by all three methods (S. Fig. 7), however there were clearly some method-dependent
193 differences in the proteins identified (discussed extensively elsewhere [39,48,49]). Transposon

194 mutagenesis screens are inherently limited by their inability to screen regions of the genome
195 that cannot viably be disrupted, such as essential genes. However, we were able to detect some
196 essential genes, such as within the *mur* operon (S. Fig. 8), where the transposon is in-frame
197 within the extreme 5' or 3' end of the protein CDS, the former due to translational read-out
198 from the chloramphenicol resistance marker maintaining downstream expression [45]. Thus,
199 detecting essential proteins is possible with transposon mutagenesis, but requires saturating
200 mutagenesis.

201 To identify putative new genes, we reviewed the remaining translation-fusion sites that
202 were either within genes but out-of-frame, suggestive of a nested gene, or intergenic and
203 therefore candidate new coding sequences. While the translation-fusion data can reveal open
204 reading frames that are translated, it cannot identify which codon is initiating translation.
205 Although the most common start codon in *E. coli* is AUG, one study found 47 (out of 64)
206 codons are capable of initiating translation, albeit at varying strengths [50]. To be conservative
207 we focused on the strongest 7 codons that can initiate translation: AUG, GUG, UUG, CUG,
208 AUA, AUU and AUC, which all displayed expression levels >100-fold higher than the control
209 cells [50]. We reviewed both the nested 'out-of-frame' and intergenic reading frames identified
210 in our translation screen for these start codons and identified a total of 215 unique putative
211 protein coding sequences (S. Table 2); 125 were nested gene candidates while 90 were
212 intergenic. Most annotated genes containing a predicted nested gene had only 1 internal nested
213 gene (S. Fig. 9), however, there were a handful with multiple predicted nested genes. Although
214 multiple nested genes within a parent gene has been previously documented in *E. coli,* such as
215 within *waaL* which contains 3 nested genes [38], another possible explanation for multiple nested
216 candidates within the TIS data is that these arise from ribosome slippage during translation of
217 the primary transcript, as *E. coli* K-12 strains are prone to higher rates of ribosome slippage

218 than other *E. coli* lineages [51], and the genes with ≥5 nested genes are some of the most highly
219 expressed proteins [47].

220 The predominant predicted start codon in the 215 candidate CDSs was AUG (43.72%)
221 followed by GUG (24.65%) and UUG (11.16%) (Fig. 3B). Although the initial aim of this
222 screen was to identify protein coding sequences irrespective of genome annotation, we
223 observed that most of the new CDSs were small proteins <50 amino acids (aa; median 25 aa,
224 smallest 1 aa, largest 138 aa; Fig. 3C). We did not apply a lower limit threshold for the size of
225 putative CDS as (1) translation of 2 codon sequences is known for *E. coli*, although these
226 sequences often represent translation-control mechanisms rather than a functional protein and
227 (2) the calculated sizes are derived from an estimated start codon and therefore might not be
228 accurate. The finding that most putative new CDSs were <50 aa is not entirely surprising and
229 is consistent with the hypothesis that small proteins are prevalent yet have been systematically
230 overlooked during bacterial genome annotation. Surprisingly, comparison of the candidate 215
231 CDSs with the 68 candidate small proteins identified by Ribo-seq methods by Weaver *et al.* [38]
232 revealed only 9 genes that were identified by both methods. Of note, many of the small proteins
233 identified uniquely by Ribo-Ret appear recalcitrant to transposition, precluding them from our
234 screen. This raises the possibility that small genes have specific regulatory mechanisms that
235 limit their disruption or represent an overlooked class of essential genes. However, many of
236 the small genes detected using both methods have additional neighbouring putative small genes
237 within our data, suggesting organised transcriptional units of small genes. For example, our
238 data is in agreement with the Ribo-Ret data for the identification of small genes *mgtT*, *mdtU*
239 and *ysgD*, but we also detected translation in alternate reading frames neighbouring these genes
240 (Fig. 3D). For *mgtT* and *mdtU*, we identified translation upstream of these genes, which
241 coincided with putative CDSs with a AUG start codon (CDS58 and CDS61 respectively), yet
242 these were not detected by Ribo-seq. Moreover, within *mdtU* we observed translation earlier

243 within the reading frame than the given start codon for *mdtU*, suggesting an alternate start
244 codon further upstream. For *ysgD*, we detected translation both within this reading frame and
245 within an overlapping reading frame, consistent with CDS177, which also has a putative start
246 codon (AUC). Interestingly, Weaver *et al.* [38] observed translation extending beyond the stop
247 codon of *ysgD*, which would be consistent with expression of CDS177, but the absence of an
248 AUG, GUG or UUG start codon filtered CDS177 from their analysis.
249
250 ***Validation of new genes***
251 To assess the validity of our method, we selected 17 putative new CDSs identified by this
252 method for verification by western blot analysis. We chose candidates with a cross-section of
253 genomic neighbourhoods (start-, mid- and end-operon), in addition to some nested genes for
254 proof of principle, and focused on these 17 because of their gene-neighbourhood with some
255 well-known and extensively studied genes (Fig. 4A). Of the 17 putative CDSs, 12 are
256 intergenic and candidate operonic genes, while the remaining 5 are nested genes but in a
257 different reading frame to the 'parent' annotated gene; 1 of the nested genes (within *damX*)
258 was only identified in one of the kanamycin-selection replicate datasets, but because of its
259 unusual transposon insertion profile we included it for follow-up investigation. Using lambda-
260 red mediated recombination, a dual epitope "sequential peptide affinity" (SPA) tag, comprising
261 a calmodulin binding peptide and 3×FLAG tags, was inserted into the genome immediately
262 upstream of the native stop codon. Constructs were checked by PCR (S. Fig. 10) and confirmed
263 by Sanger sequencing. This approach has been demonstrated previously for the detection of
264 small proteins [52,53], as the large size of the tag (~8 kDa) can enable detection of proteins as
265 small as 1 kDa. Strains were grown in LB broth to both exponential and stationary phase, cells
266 were harvested and the whole cell lysates probed with anti-FLAG antibody for the detection of
267 endogenously expressed small proteins. We detected expression for 6/12 intergenic CDS, and

268 3/5 of the nested CDSs (Fig. 4B). Small proteins in *E. coli* vary in their detection levels by
269 western blot analysis [52], and there were visible differences in the levels of protein detected here.
270 CDS58, encoded upstream of *mgtST*, had the weakest signal suggesting this protein is not stable
271 or is present at very low levels. We observed that the number of reads for each small protein
272 did not correlate with the strength of western blot detection, for example, CDS58, which is the
273 faintest on a western blot, had 93 reads, whereas CDS119 and CDS147 were very prominent
274 by western yet only represented by 9 and 12 reads respectively. As such, we did not set a
275 minimum read threshold for CDS detection provided the CDS was detected in two independent
276 kanamycin-selection screens. It's important to note that western blot detection of the C-
277 terminal FLAG tag can validate endogenous translation of a given reading frame, but cannot
278 confirm a predicted start codon for initiating translation. For example, we detected the
279 expression of tagged CDS66, however, the migration of this protein was indicative of a larger
280 protein than expected. We confirmed by Sanger sequencing that the correct reading frame was
281 targeted, and the SPA tag is out-of-frame with the *ompC* start codon. Two possible explanations
282 are (1) a homomultimeric complex that is not reduced by the denaturing conditions when
283 separating the proteins or (2) translation is initiated from the *ompC* start codon and a +1
284 frameshift enables expression of the SPA tag, however the predicted size of this construct (with
285 tag) is ~34.5 kDa.

286 Of the CDSs that we did not detect, it's possible that they are only transiently expressed,
287 are susceptible to rapid degradation, are expressed at levels below the threshold for western
288 blot detection, or are destabilised by C-terminal location of the tag. Indeed, a study validating
289 small proteins identified by Ribo-seq methods found that some could not be detected by
290 western blot analysis until the addition of a ClpP protease inhibitor, bortezomib [41]. Another
291 possibility is that some of these proteins might be secreted, and therefore missed during our

292 cell harvesting steps prior to lysing the cells for western blot analysis. However, none of the
293 putative CDSs contained any predicted signal sequences [54].
294 We next predicted the secondary and tertiary structures of the small proteins using
295 PSIPRED and AlphaFold2, including transmembrane (TM) domain prediction using
296 MEMSAT-SVM. CDS42, CDS43, CDS58, CDS61, CDS146, CDS201 and CDS57 were all
297 too short for analysis by PSIPRED (minimum size is 30 residues). All proteins that passed the
298 size threshold for PSIPRED and MEMSAT analysis were predicted to contain a single TM
299 helix, with 4 of these (CDS118, 119, 147 and 87) predicted to be pore-lining (S. Fig. 11). The
300 AlphaFold2 predictions, generated by the ColabFold pipeline, largely corroborated the
301 secondary structure predictions of PSIPRED, with many helical domains predicted for these
302 small proteins (S. Fig. 12). Although the confidence metrics (pLDDT and pTM) of these
303 predicted structures are low, this likely stems from the limited depth and diversity in the
304 multiple sequence alignments (MSA) generated by the ColabFold pipeline (S. Fig. 13). The
305 AlphaFold2 system relies on deep and diverse alignments to generate high-confidence models
306 [55,56], however, newly identified proteins may have small sequence families or few homologs in
307 the ColabFoldDB database through which the MSA are generated [55,56]. Although small proteins
308 are notoriously challenging to classify, the data from three independent structural prediction
309 tools suggest some of these small proteins may be small helical proteins and possibly
310 membrane proteins. Previous studies have reported the prevalence of helical proteins in small
311 protein screens [53], with functions ranging from toxin systems that promote persister cell
312 formation, to accessory subunits of larger membrane protein complexes [57,58].
313
314 ***Conservation of new genes***
315 Finally, we evaluated the conservation of these genes within *E. coli*. To investigate the
316 prevalence and distribution of the 17 candidate CDSs among *E. coli* strains, we screened a

317 collection of 227,873 globally distributed *E. coli* genomes (S. Fig. 14). Using tblastn analysis,
318 we identified that most of the CDS were conserved in most (>95%) *E. coli* genomes (Fig. 5A);
319 with the exceptions of CDS146 (92.8%), CDS66 (in-*ompC*; 89.8%), CDS57 (in-*marR*; 87.1%)
320 and CDS87 (in-*ypjA*, 44.8%). Among these data, most CDS had blast coverage and identity
321 scores >80%, suggesting these CDS are conserved. To establish whether the new small genes
322 are co-conserved with their gene neighbours, we reviewed whether the genetic synteny of the
323 CDSs was preserved across the different *E. coli* isolates. A 4 kb window (2 kb up- and down-
324 stream) surrounding each CDS was extracted from the *E. coli* BW25113 genome as a reference
325 and compared by blastn with the *E. coli* database to identify whether local gene
326 neighbourhoods were maintained across the genus. Most CDS were conserved across the 4 kb
327 region indicating the newly identified CDS are highly co-conserved with their gene neighbours
328 (Fig. 6). For some genes, e.g. CDS43 neighbouring *fldA*, or CDS119 neighbouring *ptsN* and
329 *yhbJ*, gene synteny was preserved along the full 4 kb region. Whereas CDS58 was more
330 strongly co-conserved with the neighbouring *mgtST* operon than neighbouring *ydeE*, and
331 CDS115 was more strongly co-conserved with *rimP* than divergently transcribed *argG*,
332 suggesting novel leader sequences for the *mgtST* and *rimP-nusA* operons, respectively. Genes
333 are often co-localised with functionally related genes and the level of conserved organisation
334 might hint at the potential functions of these newly identified genes. Altogether the new small
335 genes were co-conserved with genes with a range of biological functions, highlighting the wide
336 degree of biological potential that may be overlooked by excluding small proteins.
337

338 **Discussion**

339 The *E. coli* genome has been highly scrutinised for the identification of novel small proteins
340 [9,41,52], yet this method still revealed additional new proteins with the detection of >200 putative
341 new coding sequences. We validated the endogenous expression of small proteins detected

342 exclusively by this approach using western blot analysis, and confirmed detection of 9/17 of
343 these CDSs, three of which are encoded within overlapping genes. Many of the newly
344 identified genes were conserved throughout *E. coli*, and often co-conserved with their genetic
345 neighbours. Small proteins are increasingly being recognised for their diverse biological roles,
346 which often includes the modulation and regulation of larger protein complexes [59]. Our
347 identification and detection of small proteins co-conserved with genes involved in electron
348 transfer (*fldA*), magnesium transport (*mgtST* [60]), ribosome maturation (*rimP* [61]), spermidine
349 efflux (*mdtUJI* [62]), LPS export (*lptB* [63]), nitrogen-responsive gene regulation (*rpoN*), potassium
350 transport (*ptsN* [64]), regulatory RNA stability (*rapZ* [65]), queuosine precursor transport (*yhhQ* [66])
351 NAD+ biosynthesis (*nadR*) and tRNA methyltransferase (*trmI*), among others, suggests further
352 functional contributions of small proteins and highlights the functional importance of small
353 proteins that are still being overlooked. We present this as a high-throughput method
354 complementary to Ribosome-trapping or mass spectrometry-based approaches for detecting
355 expressed proteins.

356 Our gene neighbourhood conservation data and comparison with Ribo-seq data
357 revealed conserved clusters of small genes organised into transcriptional units of small genes,
358 e.g. CDS58-*mgtST*, CDS61-*mdtUJI*, or CDS177-*ysgD*. The detection of additional genes not
359 previously identified by Ribo-seq might be due to detection sensitivity (discussed below) but
360 was in part due to our wider acceptance of putative start codons: where Weaver *et al*. used a
361 conservative shortlist of AUG, GUG and CUG codons [38], we allowed for 7 putative start
362 codons, which enabled the positive identification of CDS177 overlapping *ysgD*. However,
363 these data are likely still an underestimate of the total proteome as earlier work using a GFP
364 reporter has shown as many as 47 of the 64 codons (including one canonical stop codon) are
365 able to initiate translation [50]. Ribo-Ret analyses that allow for a wider selection of start codons
366 observe more varied start codon usage [29], although this in part has been attributed to decreased

367 ribosome selectivity due to increased 30S subunit availability. However, consistent with the
368 hypothesis that non-canonical start codons can initiate translation, prior to filtering our data we
369 did observe translation within intergenic regions with no obvious start codon. One recent study
370 combining the analysis of stalled ribosomes at start and stop codons respectively identified
371 almost 400 putative CDSs, however they attribute this high number in part to pervasive
372 translation [41]. While pervasive translation, described as the spurious translation of a non-
373 functional coding sequences, could contribute to false positive translation-detection in our
374 screen, we observed that the detection rate of expected functional translation-fusions was quite
375 low (27.5%: the number of detected correct-RF insertions within annotated genes/the total
376 number of correct-RF insertions for those genes) suggesting active translation alone is
377 insufficient for positive translation-fusion detection, and that a degree of mRNA stability is
378 required for successful aminoglycoside phosphotransferase expression.

379 It's also important to note we don't yet know the detection limits of western blot
380 detection. For example, for two of the CDS we could not validate by western blot analysis, we
381 did detect faint bands at the expected migration site with excessive blot contrast enhancement,
382 although deemed these insufficient for validation and excluded them from our data. It is
383 possible that a transposon mutagenesis approach can offer an improved detection resolution as
384 compared to detection by standard western blot, especially for small proteins that are too
385 transiently or weakly expressed to be detected by standard western blot approaches without
386 purification. As both the transposon translation reporter and SPA-tag detection by western blot
387 inherently rely on the same principles of C-terminal tagging of endogenous proteins, an
388 exciting option for future method-development here might be the introduction of a transposon-
389 based SPA tag for whole-cell purification of small proteins.

390 One limitation of this method is that essential genes, or genes that are not easily
391 disrupted, e.g. those bound by DNA-binding proteins, will be less well-represented within the

392 transposon mutant pool, limiting their opportunities for detection. Secondly, this approach is
393 inherently biased against the detection of small proteins, as shorter genes will also be
394 represented by fewer translation-fusion mutants. Our data suggests that both of these
395 limitations may at least in part be overcome using a mini-Tn*5* system that can achieve near
396 saturated mutagenesis. Another strength of this method is its bp level of resolution for
397 identification of the precise reading frame where translation is occurring. However, as this
398 method identifies functional translation fusion sites, there is always the possibility that an out-
399 of-frame insertion is not evidence of a new gene, but evidence of ribosome slippage and a shift
400 in the translation reading frame (programmed or otherwise).

401 Besides predicting new proteins, this method can also validate whether annotated genes
402 are translated. One example of this in our data was within the 'pseudogene' *ilvG*. Curiously,
403 our data suggests both that the N-terminal fragment of the pseudogene is still expressed, and
404 that a second reading frame (+1) downstream is also expressed. We also detected translation
405 within other unexpected genomic regions, including within a handful of repetitive extragenic
406 palindromic (REP) elements, consistent with another study measuring transcriptomics and
407 proteomics in *E. coli* O157:H7 [67].

408 Although the method of coupling random mutagenesis with reporter systems is not
409 novel [44], it is still surprisingly in its infancy when applied at scale, with only one other
410 saturating mutagenesis reporter study ('ProTinSeq') published earlier this year in the model
411 system *Mycoplasma pneumoniae* [68]. As a method it is applicable to any organism that is
412 genetically tractable and the scope of application is much broader than the system presented
413 here. One obvious adaptation of this method is to replace the reporter marker with a non-lethal
414 selection marker, such as a fluorescent protein, which, coupled with fluorescence activated cell
415 sorting, would enable rapid *in vivo* screening for condition-dependent changes in expression.
416 This would expand the capabilities for small protein detection as one often cited limitation is

417 that small proteins might only be expressed transiently, at low levels, or under specific
418 conditions. With many of these tools still in their relative infancy, and focusing on a limited
419 range of growth conditions, it is likely there are still more small proteins to be discovered in
420 the *E. coli* genome. In particular, nested genes have been gaining increasing interest and
421 traction within the field [69–71]; with data presented here adding another line of evidence in
422 support of their wider existence.

423 Finally, this work highlights that a standardized approach to genome annotation is
424 needed [72,73], and this has been recognised with the implementation of the NCBI Prokaryote
425 Genome Annotation Pipeline applied to all newly submitted sequence data. However, we found
426 that even the most "up to date" annotation of *E. coli* K-12 strain BW25113 (re-annotated 24-
427 APR-2023) did not include many of the small proteins identified and validated in Ribo-seq
428 experiments from 5 years ago within the same (or closely related) strains [9]. As such, it's worth
429 noting that even for one of the best curated strains, *E. coli* K-12, many high-throughput
430 functional genetics screens that rely on genome annotations for genotype-phenotype results are
431 still overlooking small proteins.

432

433 **Materials and Methods**

434 ***Bacterial strains and culture conditions***

435 *E. coli* K-12 strain BW25113 ($rrnB_{T14}$ $\Delta lacZ_{WJ16}$ *hsdR514* $\Delta araBAD_{AH33}$ $\Delta rhaBAD_{LD78}$) was
436 routinely used [74]. Cell cultures were grown overnight in LB broth (Tryptone 10g/L, Yeast
437 extract 5g/L, Sodium Chloride 10g/L) at 37°C supplemented with 100 µg/ml carbenicillin
438 (Sigma) for plasmid selection, and 35 µg/ml chloramphenicol (Merck) or 25 µg/ml kanamycin
439 (Sigma) for transposon selection.

440

441 ***Cloning***

442 The transposon sequence was constructed by stitch PCR using the commercially available EZ-
443 Tn*5* Kan^{R} transposon (Lucigen), and a previously published Cm^{R} mini-Tn*5* transposon[45] as a
444 template and cloned into the pUC19 vector. The transposon was inserted between the *Hin*dIII
445 and *Bam*HI restriction sites, resulting in the plasmid "pUC19-KC-Tn", which contains the dual
446 selection transposon inserted in-frame in the LacZa gene. Mutations of the LacZa start codon
447 were introduced via site-directed mutagenesis. Briefly, the plasmid was amplified by PCR
448 using phusion polymerase (NEB) and overlapping forward and reverse primers that contain the
449 desired mutation. The vector template was digested by DpnI (NEB) following the supplier's
450 instructions. Newly amplified vector was transformed into competent DH5alpha cells (NEB)
451 and correct mutations were confirmed by Sanger sequencing before transferring to *E. coli* strain
452 BW25113. Oligonucleotides used in this study are listed in Supplementary Table 3.
453
454 ***Construction of a transposon library***
455 To make the transposome (transposon coupled with transposase), the pUC19-KC-Tn plasmid
456 was linearised by restriction digest using *Nde*I (New England Biolabs) following the supplier's
457 instructions. The mini-Tn*5* transposon was amplified by PCR using primers with a
458 phosphorylated 5' end and containing the inverted repeat mosaic ends for efficient Tn5
459 transposase-mediated transposition. The transposon was amplified using the phusion
460 polymerase (NEB) and the linearised plasmid as a template with the following conditions: 30
461 s 98°C; (10 s 98°C, 20 s 60°C, 1 min 72°C) x30; 5 min 72°C; 4°C hold. The PCR product was
462 purified via gel extraction (Qiagen) following the supplier's instructions. The Transposon was
463 eluted in TE buffer (pH 8.0; ThermoFisher Scientific), quantified by Qubit dsDNA HS Assay
464 kit (ThermoFisher Scientific) and adjusted to a final concentration of 100 ng/µl. To prepare the
465 transposome, a ratio of 1:2:1 of 100 ng/µl transposon DNA in TE buffer, EZ-Tn5 Transposase
466 (TNP92110; Gene Target Solutions) and 100% glycerol were incubated at RT for 1 h then

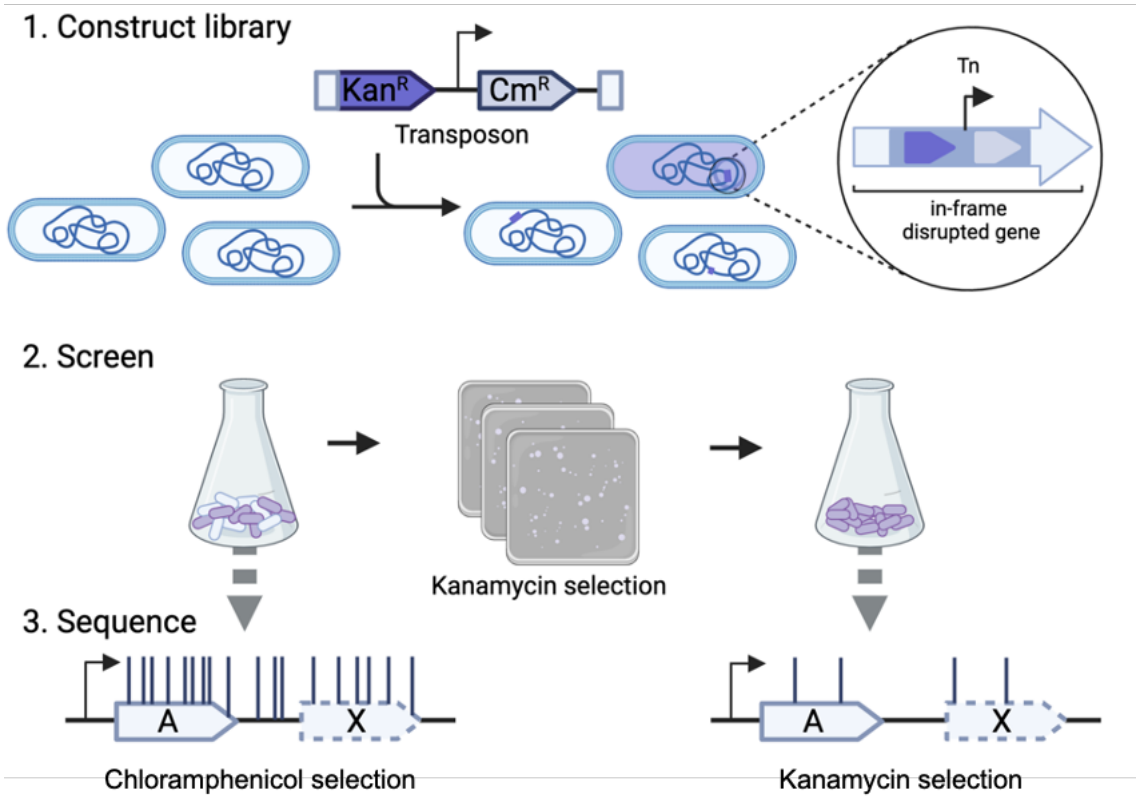
467 frozen at -20°C for storage. The transposome was introduced into *E. coli* K-12 strain BW25113
468 by electrotransformation. Briefly, cells were grown to mid-exponential phase in 800 ml LB at
469 37°C with aeration, harvested by centrifugation at 2000 xg for 10 min at RT and washed 4x in
470 10% glycerol at the same volume. The final cell pellet was resuspended in 1 ml of 10%
471 glycerol. 1 µl transposome was incubated with 100 µl cells before shocking at 25 kV. 1 ml LB
472 was added per reaction, and cells were incubated a 37°C with shaking for 2 h for recovery.
473 Successful transformants were selected on LB agar plates supplemented with 35 µg/ml
474 chloramphenicol and incubated overnight at 37°C for 18 h. Colonies were scraped from plates
475 to form the transposon library of pooled mutants and resuspended in 30% glycerol in LB for
476 storage at -80°C. For selection of kanamycin resistant mutants, 25 µg/ml of kanamycin was
477 identified as the minimum concentration of kanamycin that inhibits growth of *E. coli* BW25113
478 on supplemented LB agar plates (S. Fig. 5A) and was therefore selected as the concentration
479 for selection of kanamycin-resistant isolates. For identification of mutants expressing the
480 translation reporter, the transposon library was selected on LB agar plates supplemented with
481 25 µg/ml kanamycin, to achieve a density of ~2,000 colonies per plate. A total of ~$1x10^7$ CFUs,
482 of the input library were screened in total (per replicate), ensuring sufficient coverage of the
483 library. Colonies were scraped from plates, resuspended in 30% glycerol-LB, pooled and stored
484 at -80°C.
485
486 ***Transposon sequencing***
487 Genomic DNA was extracted, fragmented by ultrasonication to an average size of ~250 bp and
488 prepared for sequencing using the NEB Next Ultra I kit (E7370L, NEB), with some
489 modifications. Following end repair, adapter ligation and a size-selection purification step with
490 SPRI beads (Beckman Coulter) following the kit instructions, a PCR step was introduced to
491 enrich for Transposon-gDNA junctions, using a forward primer specific for the transposon and

492 a reverse primer specific for the adapter. 15 µl DNA fragments in sterile, nuclease free water,
493 25 µl Q5 polymerase, 2.5 µl 10 µM each primer and 5 µl nuclease free water to a final volume
494 of 50 µl were mixed and amplified in a thermocycler with the following profile: 3 min 98°C;
495 (15 s 98°C, 30 s 65°C, 30 s 72°C) x10; 1 min 72°C; hold 4°C. The sample was purified using
496 SPRI beads at a ratio of 0.9:1 beads to sample. A second PCR step then prepared the
497 transposon-gDNA junctions for sequencing using custom forward primers specific for the
498 transposon, which incorporated a barcode for sample identification as well as staggering the
499 start of the transposon sequence during sequencing. The reverse primers, NEBNext Multiplex
500 Oligos for Illumina (E7335, NEB), have homology to the adapter. The purpose of this step is
501 to introduce barcode sequences and further adapters to enable fragment binding to the
502 sequencing flow cell; the fragments were prepared as before with the following conditions: 3
503 min 98°C; (15 s 98°C, 30 s 65°C, 30 s 72°C) x20; 1 min 72°C; hold 4°C. The final product was
504 purified again using SPRI beads and then quantified by qPCR following the kit instructions
505 (KAPA Library Quantification Kit, Roche). Fragments were sequenced using an Illumina
506 MiSeq using v3 150 cycle cartridges (MS-102-3001, Illumina), with 5% PhiX loading control
507 (Illumina).
508
509 ***Data analysis***
510 Sequencing data were processed using a previously described approach [75]. Briefly, FASTQ
511 data were first demultiplexed by the internal barcodes (introduced by the forward primer of
512 PCR2) using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html).
513 Samples were then checked for the transposon sequence in 2 steps allowing for a total of 4 bp
514 mismatches. Following identification of transposon-containing reads, the transposon sequence
515 was trimmed and the remaining read mapped using bwa mem (https://bio-
516 bwa.sourceforge.net/bwa.shtml#13), to the *E. coli* BW25113 reference genome (available from

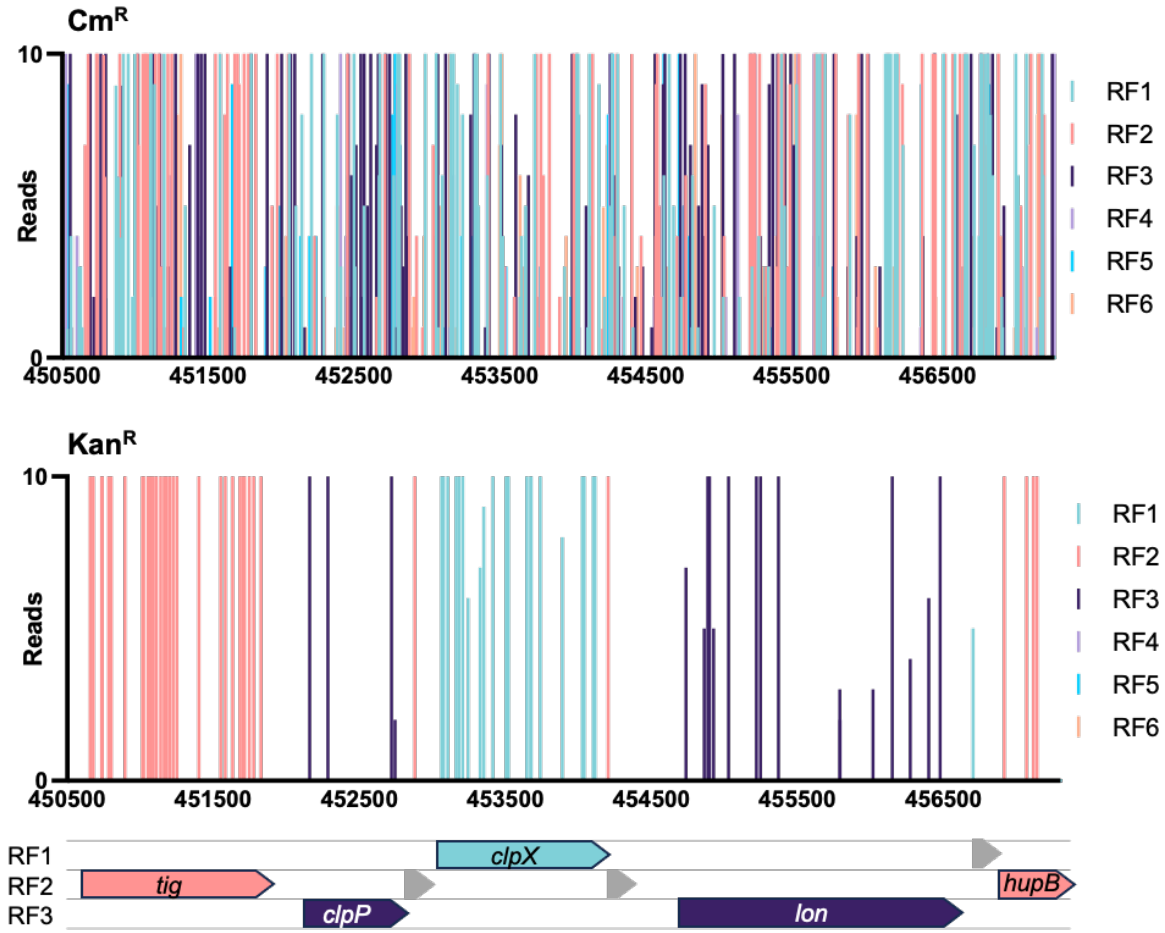
517 the NCBI database, accession CP009273.1, unless otherwise stated). The first nucleotide of a
518 given read was used as the transposon insertion site to generate insertion plot files, which were
519 viewed in Artemis [76]. Insertion data were cross-referenced with annotated genome features
520 using intersectBed [77]. We used bbcountunique.sh of BBTools for sub-sampling of FASTQ data
521 (https://sourceforge.net/projects/bbmap/).
522
523 ***Construction of SPA-tagged strains***
524 A sequential peptide affinity (SPA) tag (containing the calmodulin binding peptide (CBP) and
525 3×FLAG sequences separated by a TEV Protease cleavage site) was introduced in-frame into
526 the chromosome immediately upstream of the target CDS stop codon, resulting in a C-terminal
527 tag [53,78]; a tag routinely used for the identification of small proteins [38,53,79]. The SPA tag was
528 synthesised by GenScript and stitched by PCR [80], with a kanamycin resistance cassette flanked
529 by two FRT sites amplified from the *clsA* Keio mutant [81]. The SPA-Kan^R fragment was
530 subcloned into vector pJET1.2 (ThermoFisher Scientific), and the vector used as a template for
531 PCR amplification of the SPA-tag and kanamycin resistance cassette using primers with 50 bp
532 flanking regions homologous to the intended chromosomal insertion site. PCR profile: 98ºC 1
533 min; (98ºC 10 s, 62ºC 10 s, 72ºC 2 min)x34; 72ºC 2 min. Gel purified linear fragments were
534 electroporated into BW25114/pKD46 via the Datsenko-Wanner method of chromosomal
535 mutation [74]. All mutants were verified first by PCR and Sanger sequencing.
536
537 ***Western blot analysis***
538 Immunoblot assays to determine small protein levels were conducted as described previously
539 with minor modification [53]. In brief, overnight cultures were sub-cultured 1:500 into LB and
540 grown at 37°C with aeration to either mid-exponential phase (OD_{600} = 0.3-0.6) or stationary
541 phase (OD_{600} = 4.0-5.0). Cells were harvested by centrifugation and resuspended in 50 mM

542 sodium phosphate buffer (pH 8). Re-suspended whole cells were mixed with 4× sample buffer
543 (1× stacking buffer, 2% SDS, 0.025 mg bromophenol blue, 52% glycerol), heated at 95°C for
544 10 min and centrifuged for 10 min. Samples were separated on Novex 16% Tricine gels
545 (ThermoFisher Scientific) and transferred to nitrocellulose membranes (ThermoFisher
546 Scientific). The membranes were blocked with 2% milk and then probed with anti-FLAG M2-
547 HRP monoclonal antibody (A8592, MilliporeSigma) in PBS-T. Signals were visualized using
548 SuperSignal West Dura Extended Duration Substrate (ThermoFisher Scientific). Equal loading
549 of lanes in each gel was confirmed by examining the relative intensity of background bands in
550 the test samples as compared to the untagged, wild-type control samples.
551
552 ***Protein structure predictions***
553 Protein secondary structure were predicted using PSIPRED (4.0) with MEMSAT-SVM for
554 TM-domain prediction[82,83]. Protein structures were predicted using ColabFold (v1.5.5):
555 AlphaFold2 using MMseqs2[55].
556
557 ***Conservation analysis***
558 To investigate the prevalence and distribution of short open reading frames (sORFs) among
559 *Escherichia coli* strains, we screened a collection of globally distributed *E. coli* genomes. We
560 downloaded 227,873 *E. coli* genome assemblies, along with all available metadata, from
561 Enterobase (accessed on 14 March 2024)[84]. The presence of CDSs was determined using
562 tBLASTn (version 2.12.0+). The tBLASTn analysis was conducted with default settings, but
563 adjustments were made to account for the short length of the query sequences: -matrix PAM30
564 and -word_size 5.

565 **Figures**
566



567
568 **Figure 1. Method overview.** (1) Construction of a transposon mutant library via introduction
569 of a transposon with a dual selection mechanism. Successful transformants are isolated via
570 selection on LB agar supplemented with chloramphenicol. (2) Screening of the transposon
571 library on LB agar supplemented with kanamycin enriches for mutants that contain a
572 transposon inserted in-frame in an expressed protein coding sequence. (3) Sequencing of the
573 input and output transposon mutant pools reveals fusion mutants that resulted in the expression
574 of the kanamycin resistance cassette, and therefore identifies protein coding sequences.



575
576
577 **Figure 2. The bp resolution of the translation reporter reveals the reading frame of**
578 **protein coding sequences.** The initial construction of the transposon library was selected on
579 agar plates supplemented with chloramphenicol (top panel). The vertical bars represent both
580 the frequency and position of identified transposon insertion sites. The data are coloured
581 according to the reading frame (RF) at the site of insertion. A secondary selection on agar plates
582 supplemented with kanamycin selected for mutants where the transposon is inserted in frame
583 of a protein coding sequence. The data are coloured according to the reading frame at the point
584 of insertion and are consistent with the annotated genes (bottom panel). Annotated genes are
585 indicated by black arrows while putative new protein coding genes are indicated by grey
586 arrows.

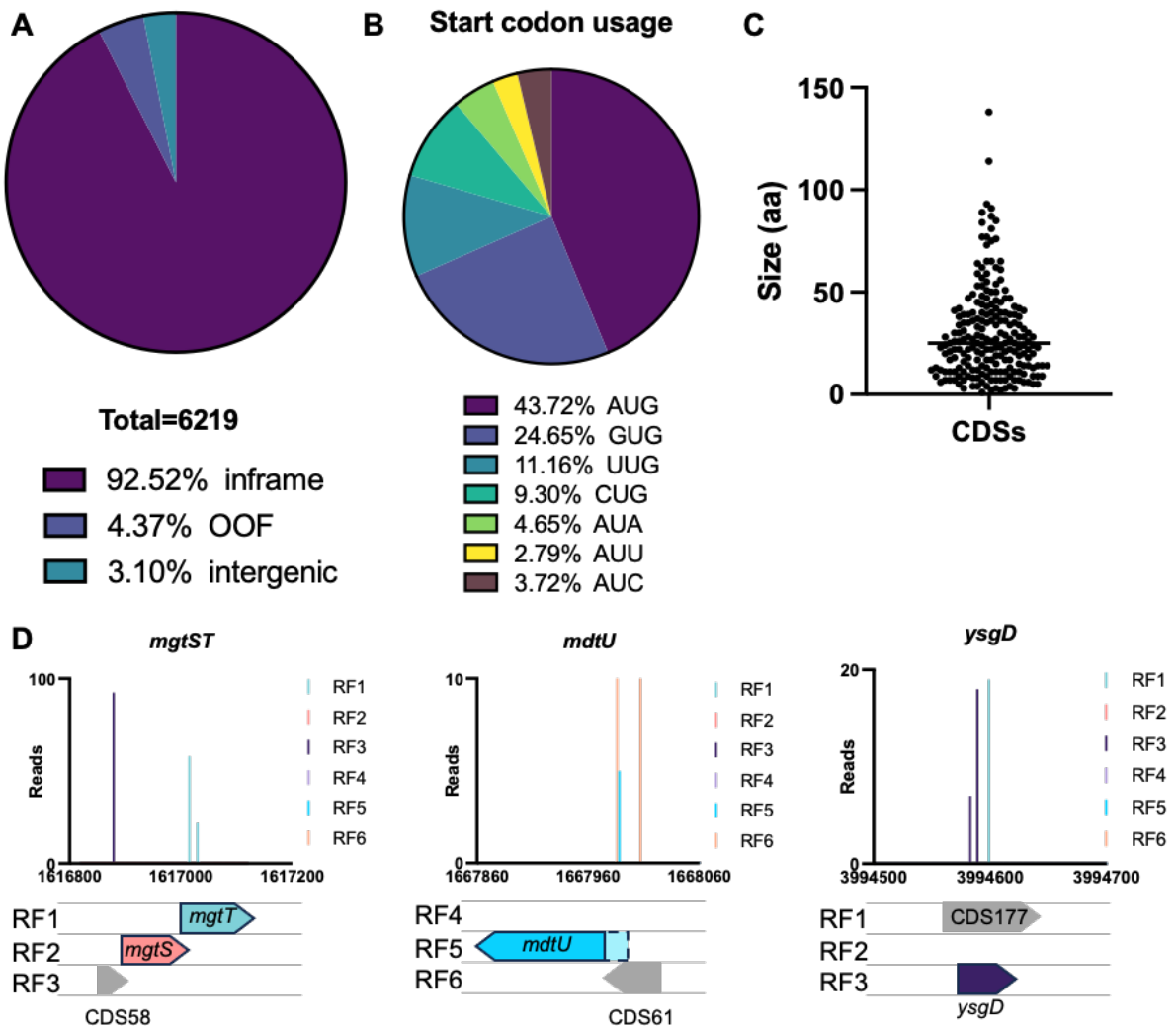


Figure 3. Identification of protein coding sequences. (A) The identified transposon-insertion site (and therefore translation fusion junction) was mapped to the reference genome and cross-references with annotation information. 92.52% of insertions were within annotated genes and 'inframe' consistent with the reading frame of the annotated gene. 4.37% of insertions were within annotated genes but in a different reading frame to the annotated gene ("out of frame [OOF]) and 3.10% of insertions were not within annotated protein coding sequences. (B)(i) The start codon frequency of 215 putative CDSs, and (ii) the CDS sizes in amino acids, with the median shown by a horizontal line. (C) Small CDSs identified by reporter TIS neighbouring small genes identified by Ribo-Ret[38]. The transposon insertion sites are those identified following selection with kanamycin (representing translation-fusion events) and are coloured according to the reading frame (RF) at the site of translation-fusion. Putative new genes are shown in grey.

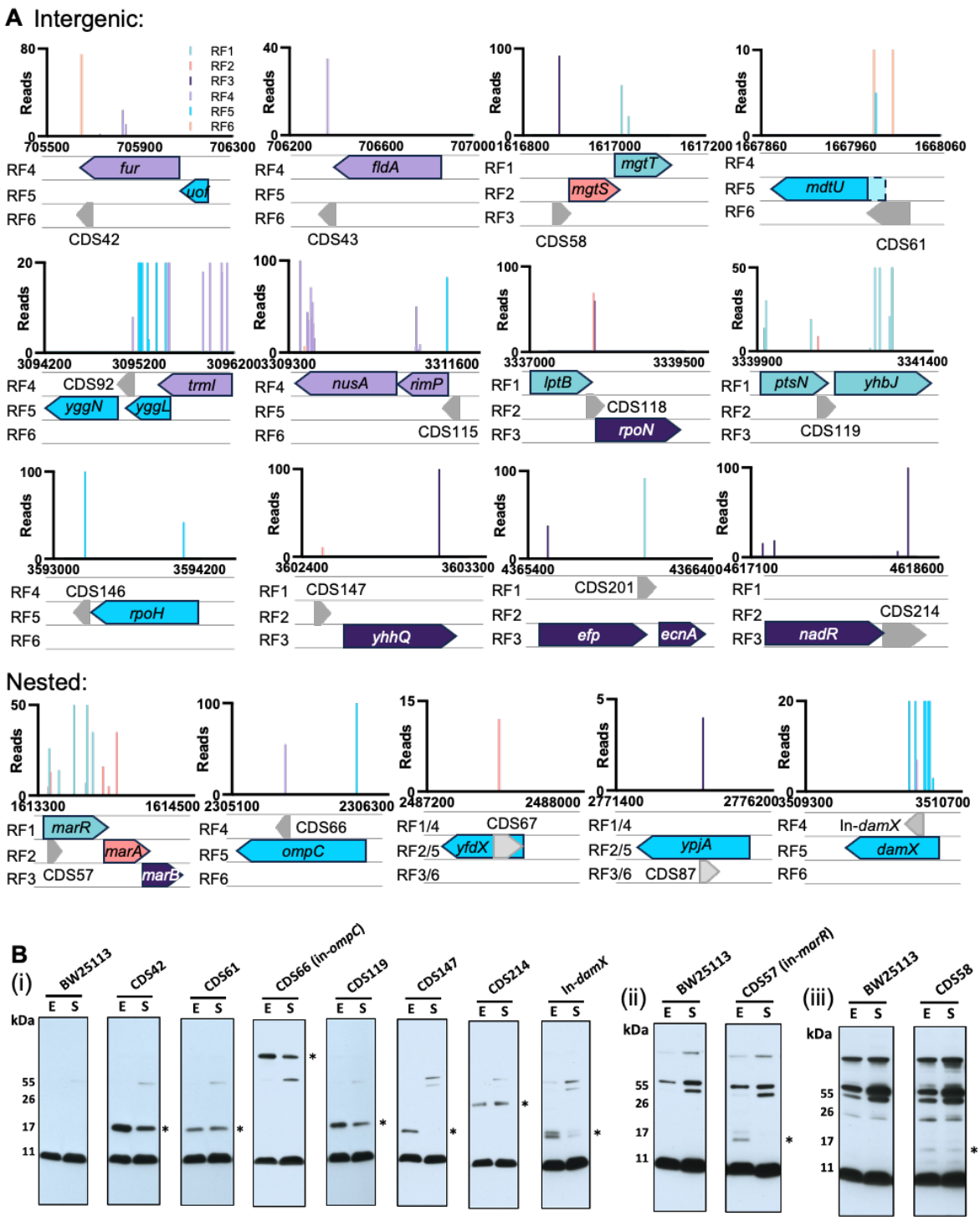


Figure 4. Validation of new proteins. (A) The genetic neighbourhoods of putative protein coding sequences (CDS) identified by reporter transposon-insertion sequencing selected for validation. Putative CDSs are shown in grey. (B) Representative western blots (of 3 repeats): whole cell lysates probed with anti-FLAG antibody, detected proteins are indicated by *. BW=BW25113 control (untagged); E = Exponential phase; S = Stationary phase. Blots are separated into three panels according to the exposure time used for protein detection (i) short (ii) medium (iii) overnight.

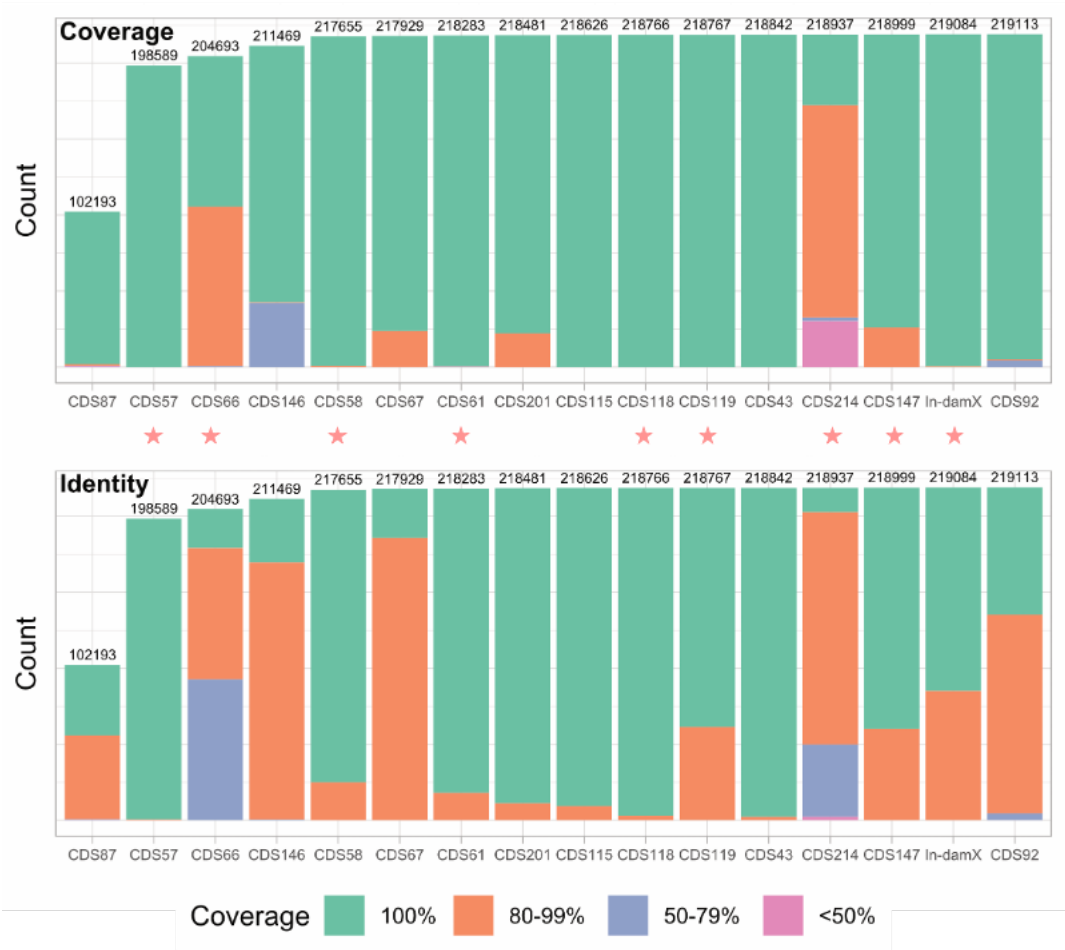


Figure 5. CDS carriage within the 227,873 genome database. tblastn results of CDS homologs within *E. coli* gnomes, coloured by % coverage and % identity accordingly. CDS validated by western blot analysis are indicated with a star.

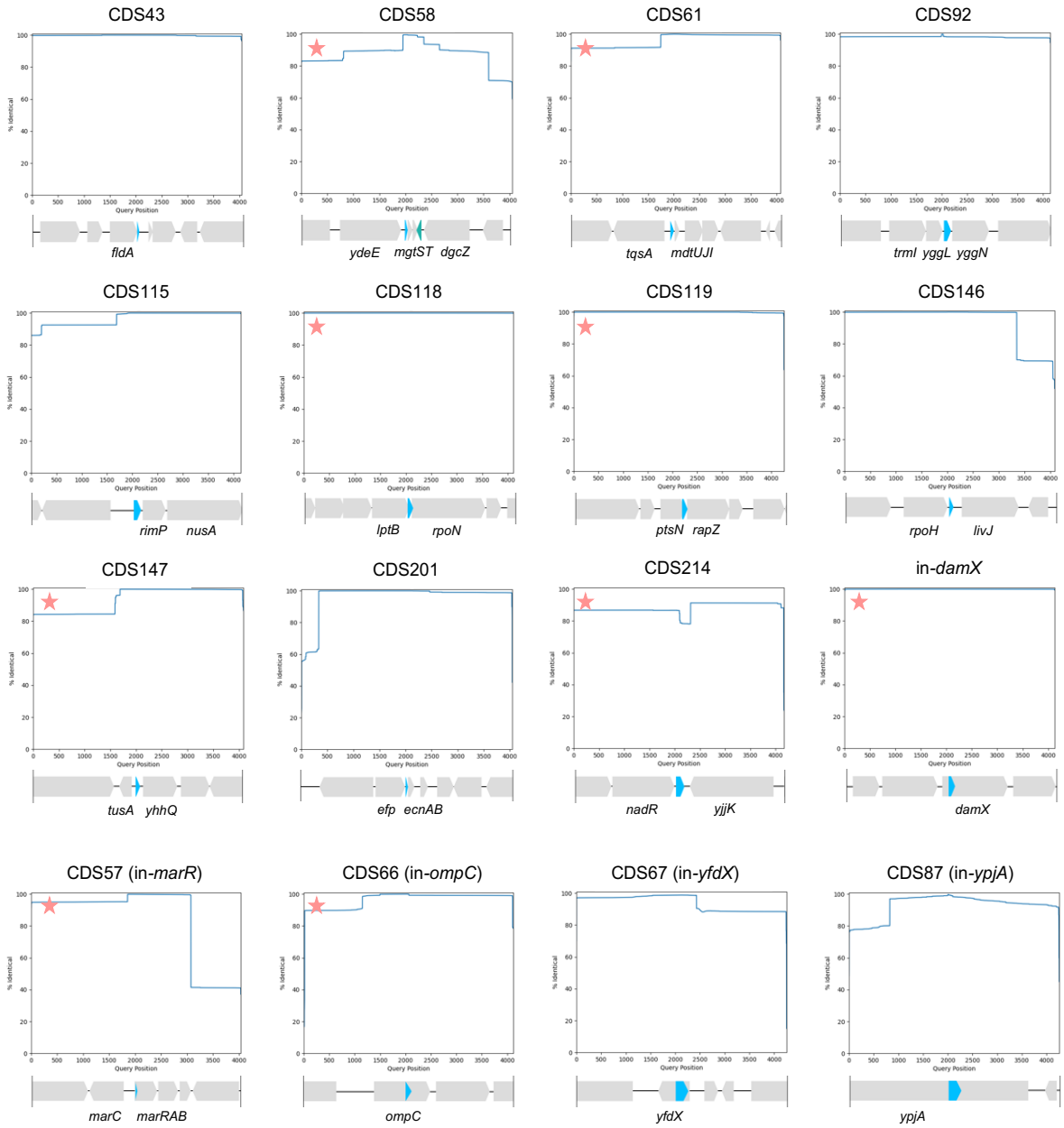
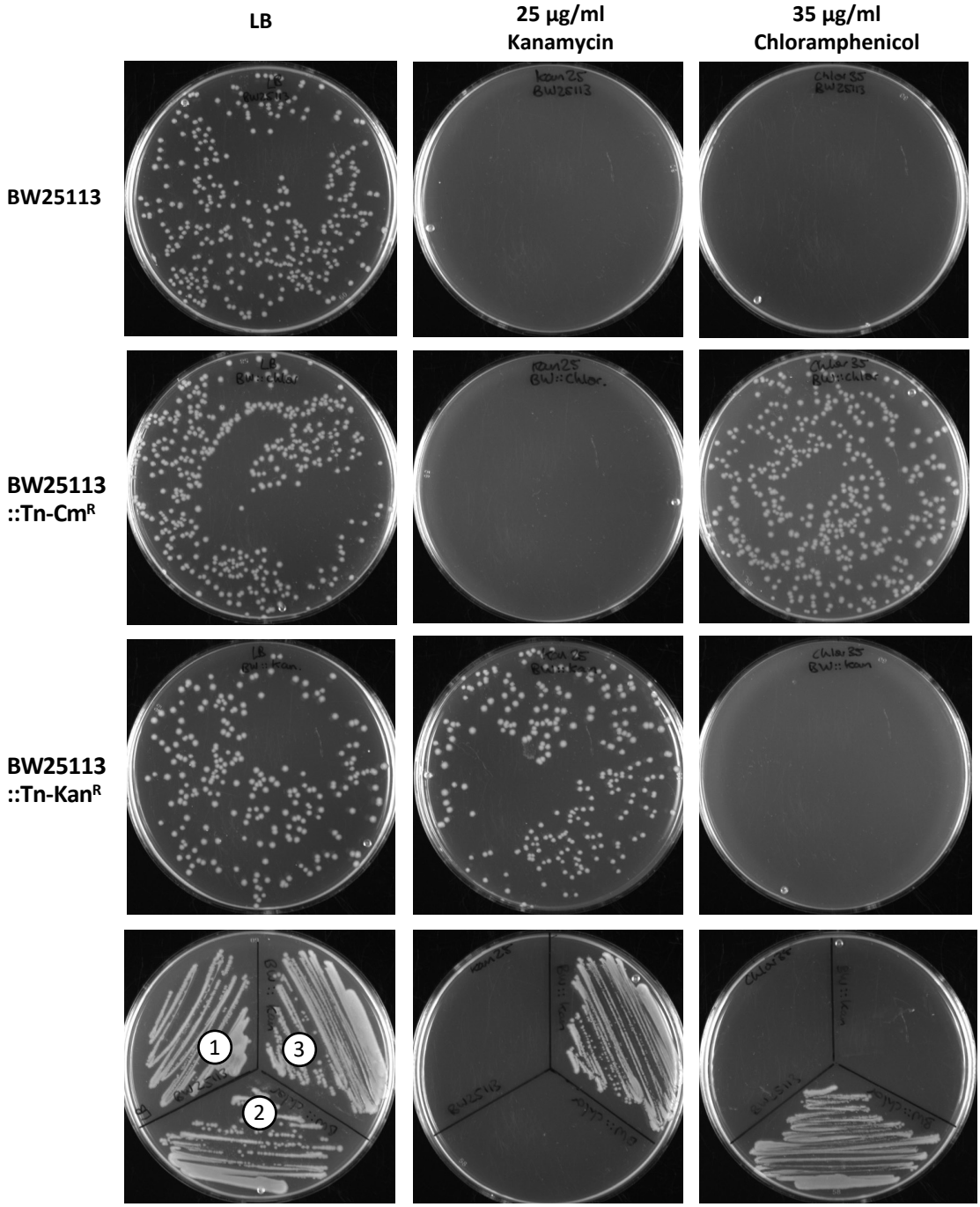


Figure 6. Blastn %identity plots of gene neighbourhoods

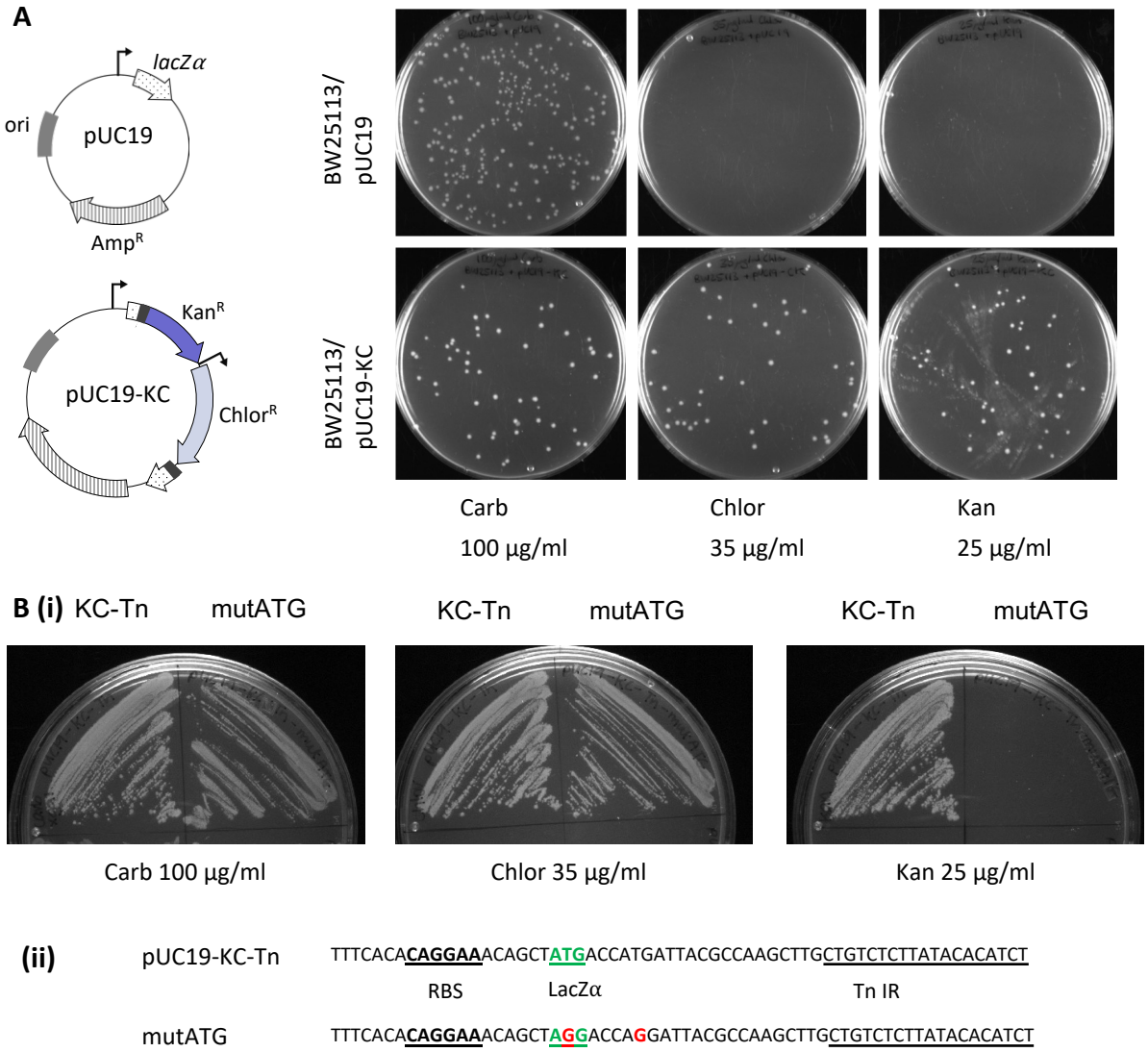
Nucleotide sequence similarity of 2 kb up- and down-stream of each CDS from the *E. coli* BW25113 reference genome compared with the *E. coli* genomes database. Each plot shows the % of genomes (y-axis) that have coverage along the reference genome query sequence (x-axis) with the BW25113 reference sequence indicated below. Annotated genes are shown in grey, candidate new genes are highlighted in blue, the *mgrR* RNA (downstream *mgtST*) is coloured in green. CDS additionally detected by western blot are indicated with a star.

622 **Supplementary Tables**
623
624 **Supplementary Table 1. Comparison between datasets**
625 **Supplementary Table 2. Putative protein coding sequences**
626 **Supplementary Table 3. Oligonucleotides used in this study**

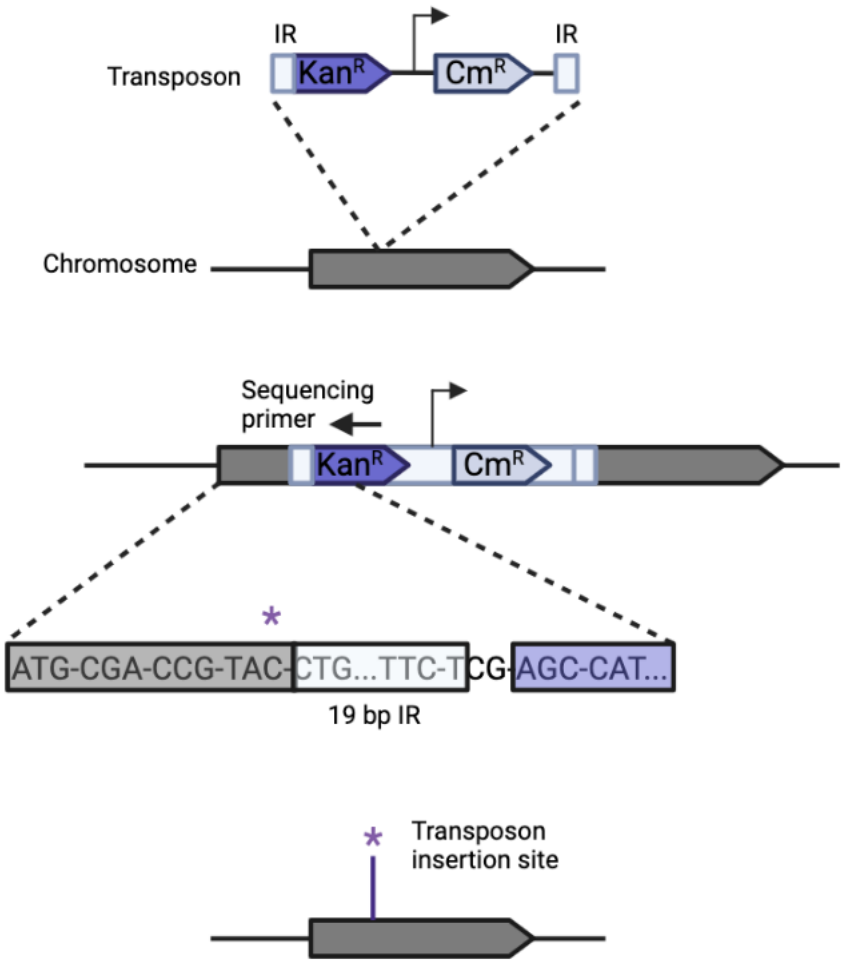
627 **Supplementary Figures**
628



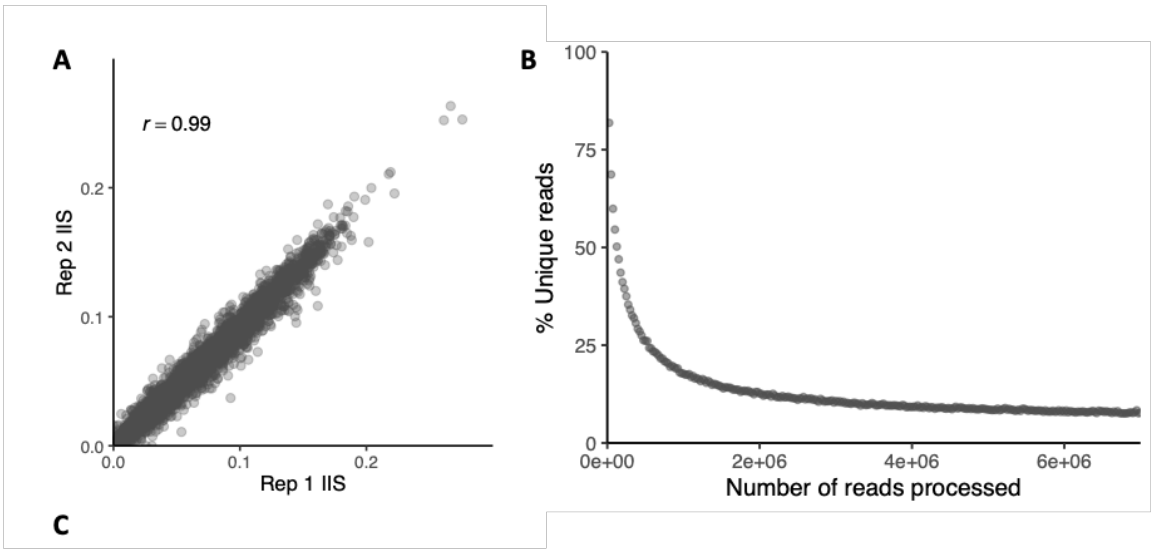
629
630 **Supplementary Figure 1. Antibiotic cassette cross-resistance**
631 Growth of *E. coli* BW25113 or *E. coli* BW25113 mutagenized at random with a mini-Tn*5*
632 transposon carrying either a kanamycin (KanR) or chloramphenicol (CmR) resistance cassette
633 on LB agar plates supplemented with antibiotic. The bottom row shows the same selection
634 plates with all strains inoculated: (1) *E. coli* strain BW25113, (2) BW25113::Tn-CmR, (3)
635 BW25113::Tn-Kan^{R}. These transposons formed the basis of the dual selection transposon,
636 stitched together by PCR.
637



Supplementary Figure 2. Translational fusion and expression of the kanamycin resistance gene.
(A) *E. coli* BW25113 was transformed with either the pUC19 plasmid or pUC19-KC, which contained the kanamycin-chloramphenicol dual selection transposon cloned in-frame into the LacZa fragment. Transformants were selected on LB agar supplemented with Carbenicillin, Chloramphenicol or Kanamycin. (B) Single nucleotide polymorphisms were introduced by site directed mutagenesis into the start codon(s) of LacZα, replacing ATG with AGG, to make the pUC19-KC-Tn-mutATG vector. *E. coli* BW25113 with either pUC19-KC-Tn or pUC19-KC-Tn-mutATG were inoculated onto LB agar plates supplemented with Carbenicillin, Chloramphenicol or Kanamycin.

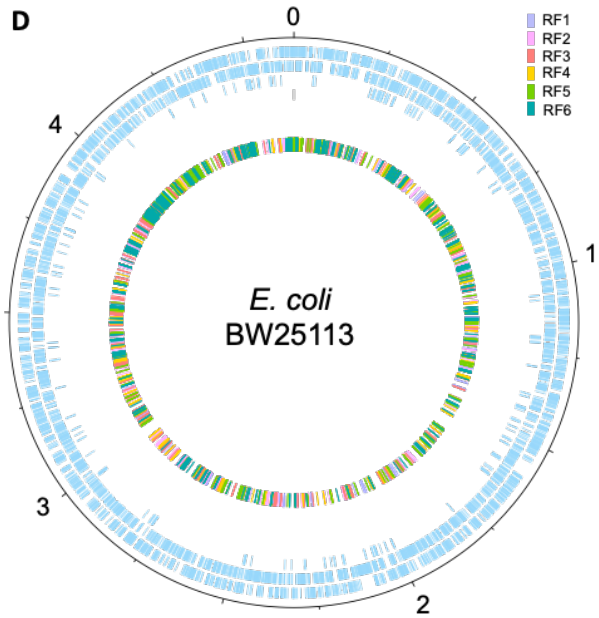


Supplementary Figure 3. Schematic for identification of the transposon insertion site
The transposon inserts into the genome at random. For identification of the transposon insertion site, the transposon-genomic DNA junction is sequenced by sequencing out from the kanamycin resistance gene into the neighboring genomic DNA. The nucleotide immediately adjacent to the 19 bp inverted repeat of the transposon (indicated by an asterisk) is identified as the transposon insertion site. Abbreviations: IR, inverted repeat; Kan^R, kanamycin resistance; Cm^R, chloramphenicol resistance.

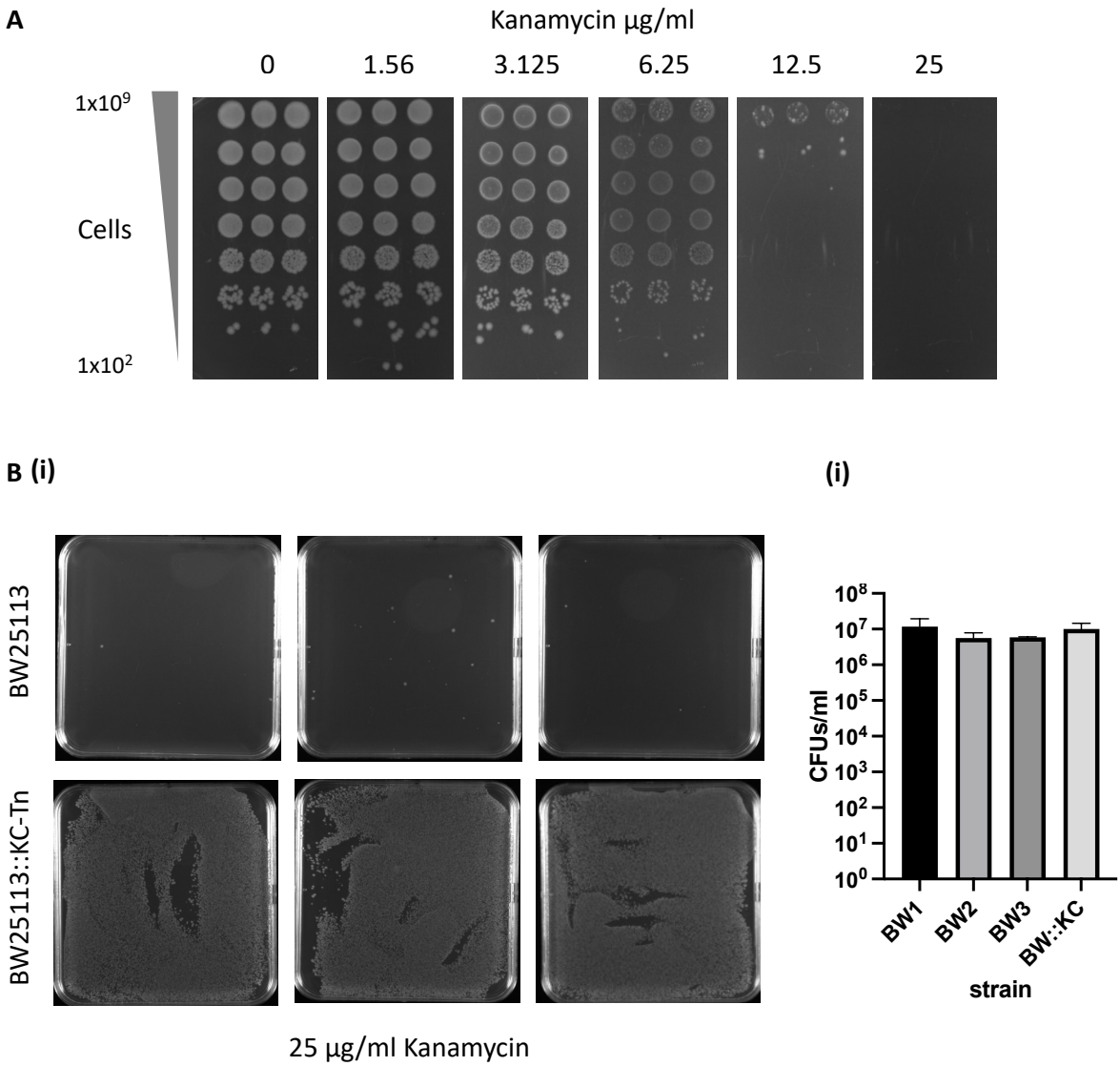


C

Reading Frame	Insertions	Percentage
RF1	61,892	16.72%
RF2	62,034	16.76%
RF3	61,783	16.69%
RF4	61,651	16.66%
RF5	61,219	16.54%
RF6	61,535	16.63%

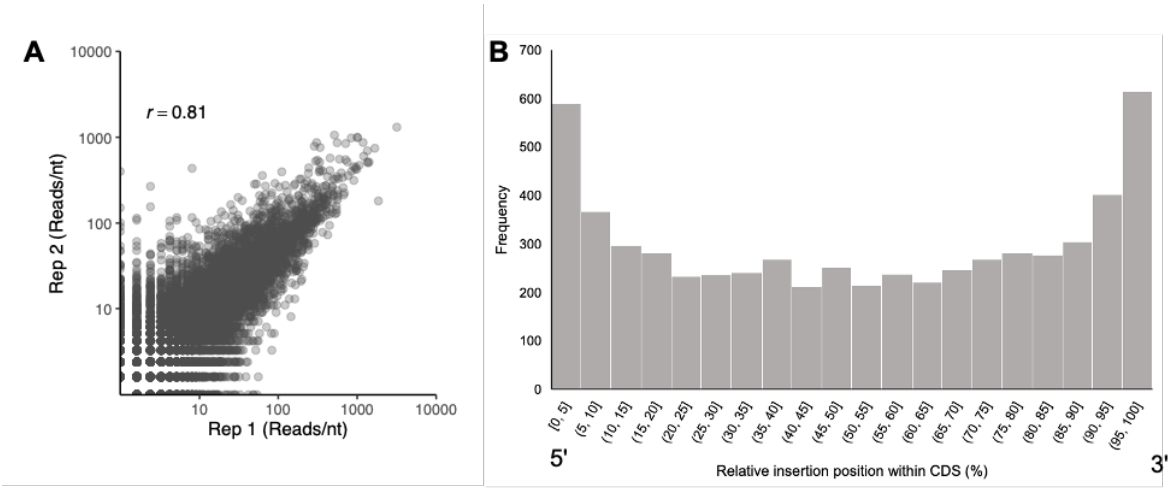


Supplementary Figure 4. Analysis of the reporter transposon library
(A) Comparison of insertion index scores (IIS; the number of insertions per gene normalised by gene length) between sequenced replicates of the input transposon library. (B) Rarefaction plot of the input transposon library. Sub-sampling of transposon-trimmed fastq data, using BBTools, shows the percentage of unique reads with increasing sampling of sequencing data. (C) Proportion of insertions per reading frame (RF) throughout the genome. (D) Chromosomal position of each identified insertion site around the *E. coli* BW25113 chromosome, and therefore translation reporter position, coloured by reading frame (RF). The two outermost tracks in light blue correspond with sense and antisense CDS respectively, with pseudogenes highlighted in light blue inside.

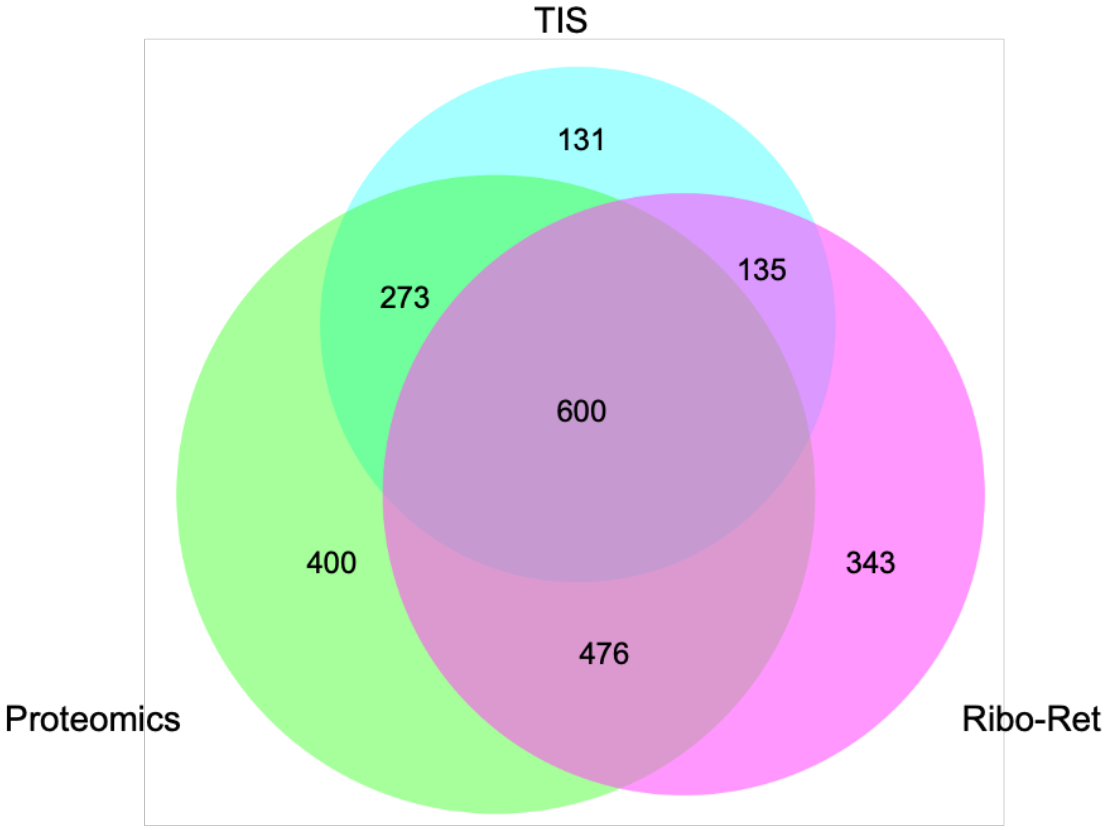


Supplementary Figure 5. Antibiotic selection screening

(A) Identification of a suitable concentration of kanamycin for agar plate selection. *E. coli* BW25113 was grown in triplicate overnight at 37˚C. Samples were normalized to an OD_{600} = 1.00, equivalent to $1x10^9$ cells. Cells were 10-fold serially diluted in LB broth and 5 µl of each dilution was inoculated onto LB agar plates with and without kanamycin at 2-fold decreasing concentrations. Concentrations of 50 and 100 µg/ml kanamycin not shown. Plates were incubated overnight at 37˚C and imaged the following day. (B)(i) Cell cultures of *E. coli* BW25113 (in triplicate) and *E. coli* BW25113::kan^R-Cm^R-Tn (KC-Tn) were normalized to an OD_{600} = 0.01 in LB; 500 µl of this cell culture was inoculated onto square LB agar plates supplemented with 25 µg/ml kanamycin and grown overnight at 37˚C. (ii) In addition, 10-fold serial dilutions of these cultures were inoculated on LB only plates to enumerate the input colony forming units (CFUs) for each sample.

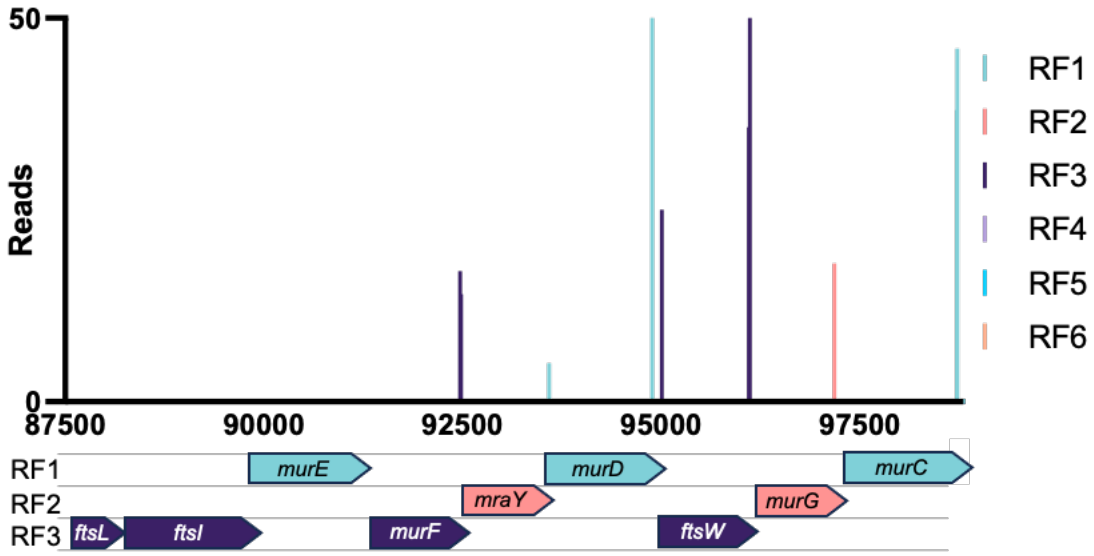


Supplementary Figure 6. Analysis of kanamycin selection mutants
(A) Comparison of the number of reads per nucleotide between replicates of the BW25113::KC-Tn library plated on LB supplemented with kanamycin. (B) The position of each insertion within a protein coding sequence relative to the gene start and end taken as a percentage.



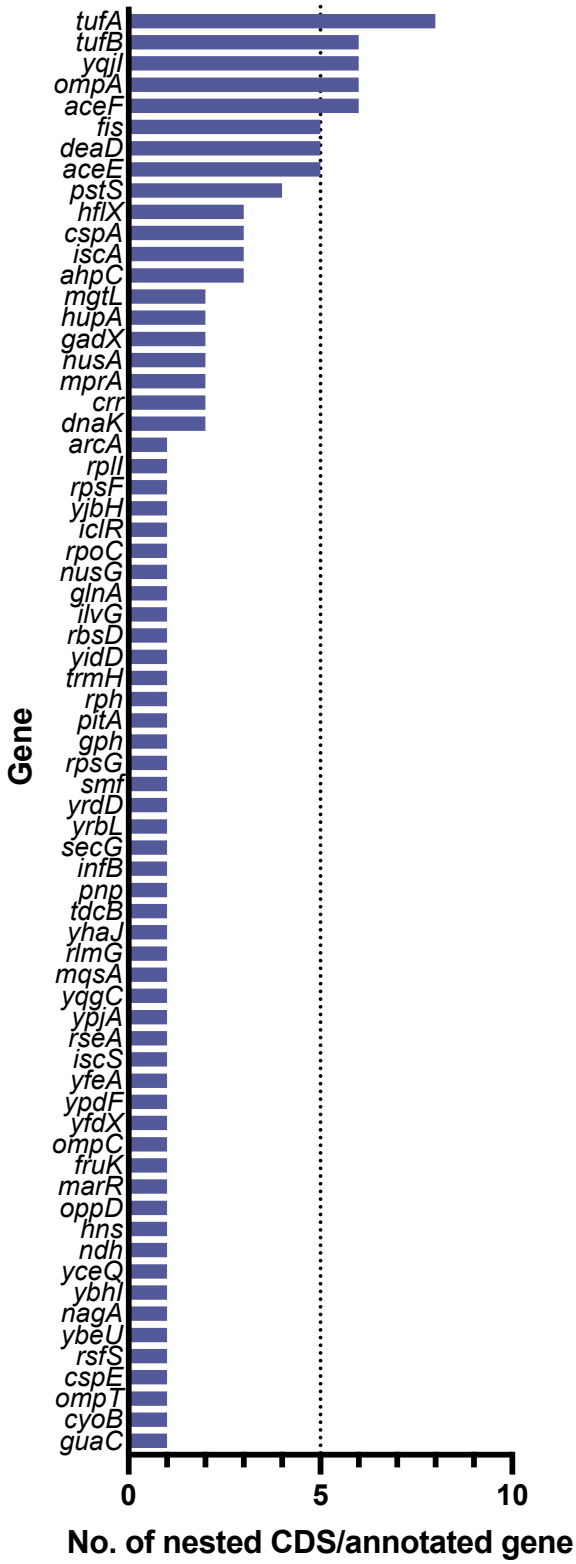
Supplementary Figure 7. Dataset comparison

Comparison of the annotated genes positively identified by different high-throughput methods for the identification of proteins/protein-coding genes on a whole cell scale. All datasets are reported for the *E. coli* K-12 stain BW25113 grown in LB, with the caveat that the transposon-insertion sequencing (TIS) data was exposed to kanamycin, while the Ribo-Ret dataset was a BW25113 derivative with the genotype BW25113Δ*tolC* and grown in LB supplemented with 0.2% glucose prior to exposure to 12.5 µg/ml Retapamulin. TIS method with a Ribo-Ret dataset (Meydan *et al.* 2019) and a Proteomics dataset (Schmidt *et al.* 2016).

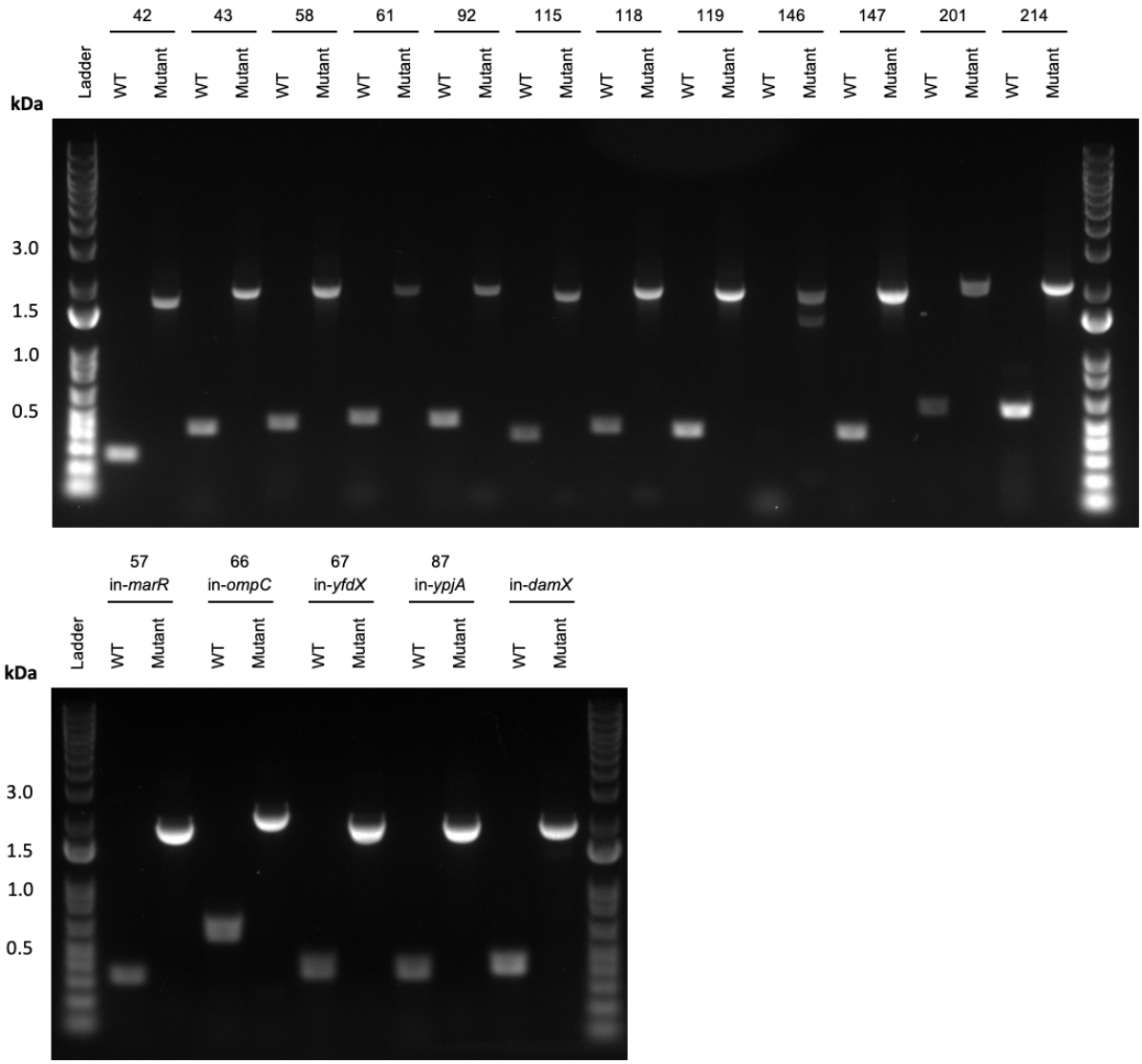


Supplementary Figure 8. Expression of essential genes detected by the translation-reporter transposon

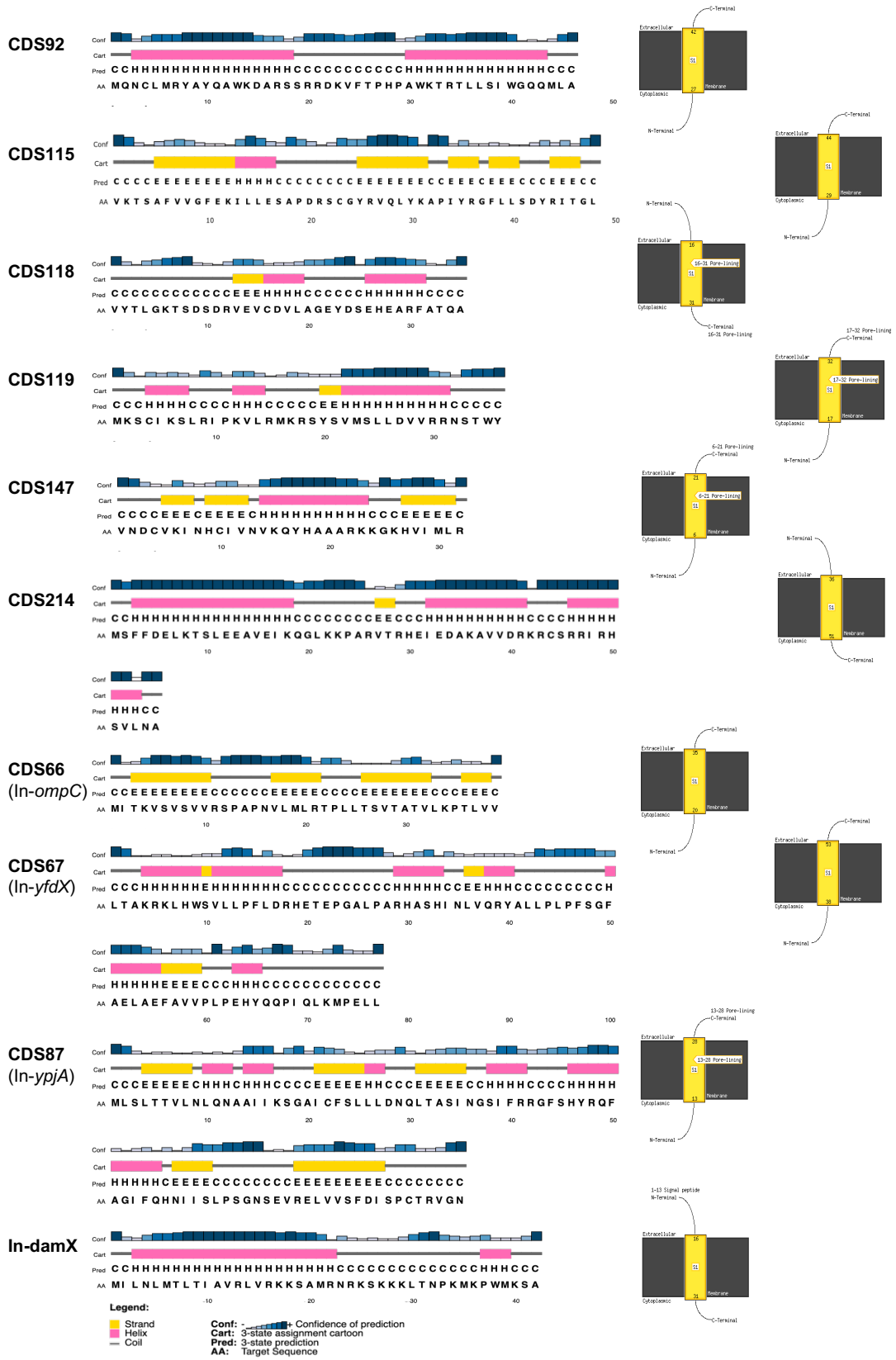
Viable translation-fusion events within 5/9 (*murF, murD, ftsW, murG* and *murC*) essential genes of the mur operon following selection of the transposon library on kanamycin. Insertions are at the extreme 5' end of each CDS, coloured according the reading-frame (RF). The presence of 5' transposon insertion events suggests translational read-out from the transposon to maintain expression of downstream essential genes.



Supplementary Figure 9. Frequency of out-of-frame nested genes within annotated genes

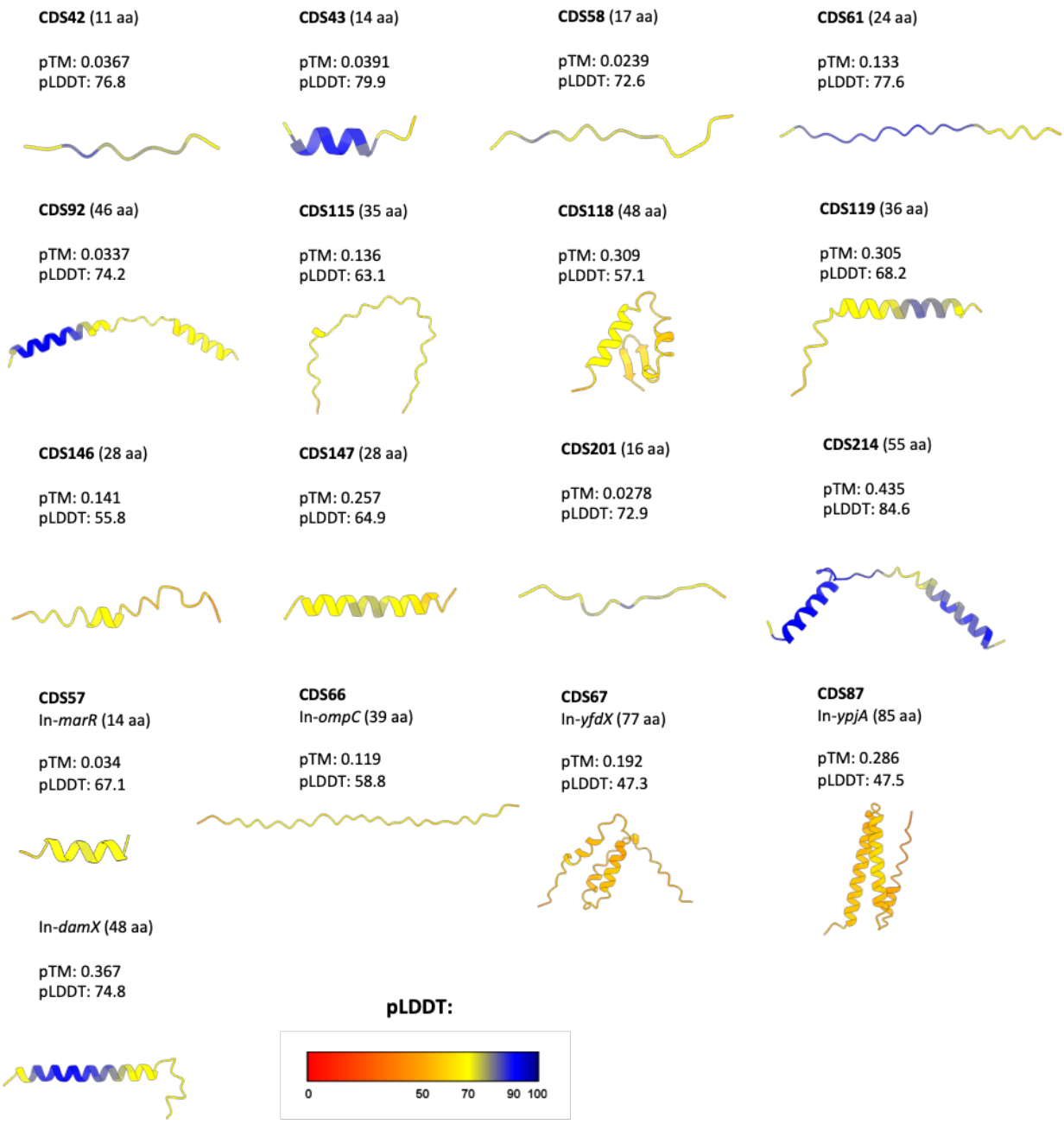


Supplementary Figure 10. Construction of SPA-tagged mutants.
PCR amplification of stop codon/fusion junction in the parent strain (BW25113) and each SPA-tagged mutant.

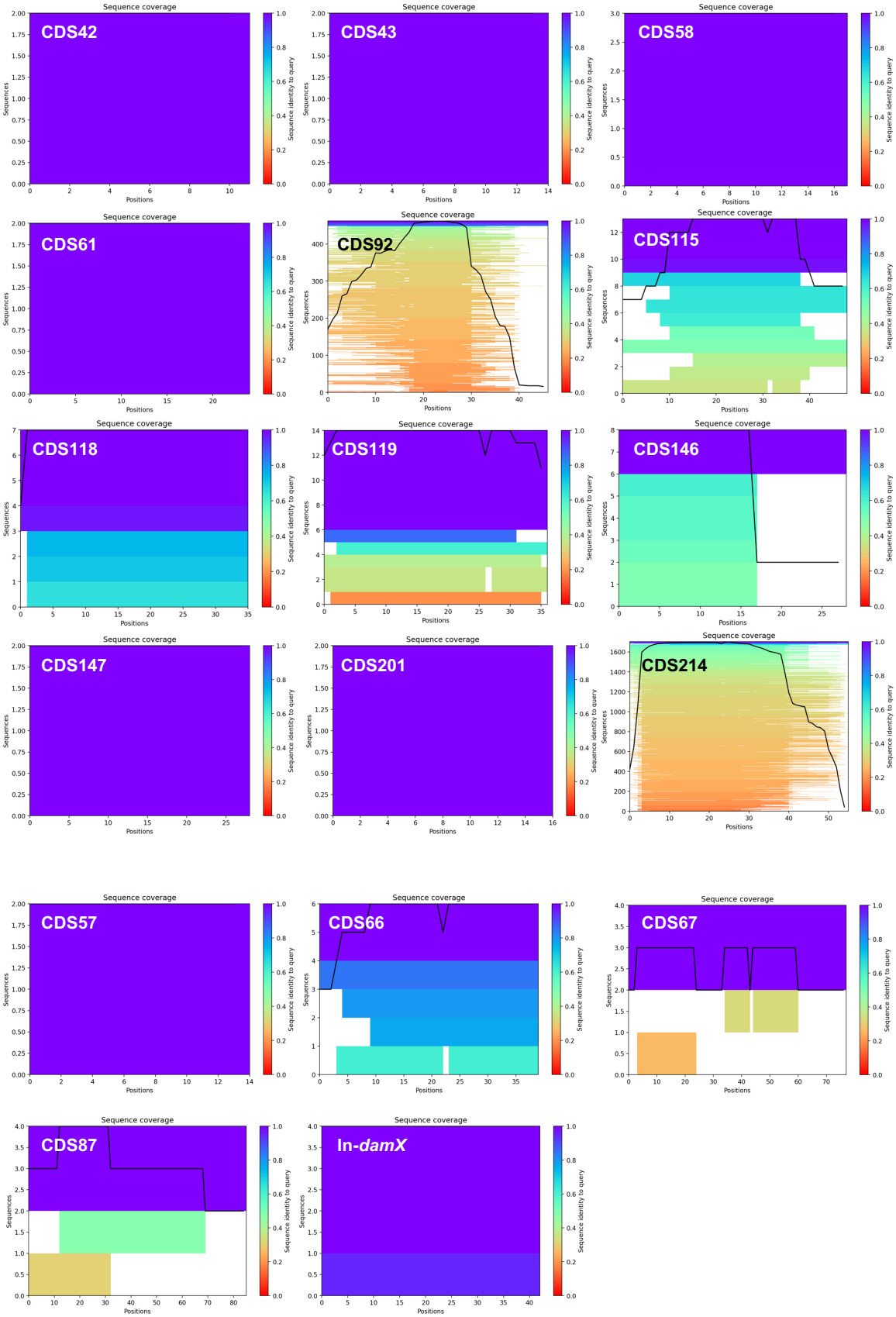


Supplementary Figure 11. Predicted protein structure using PSIPRED

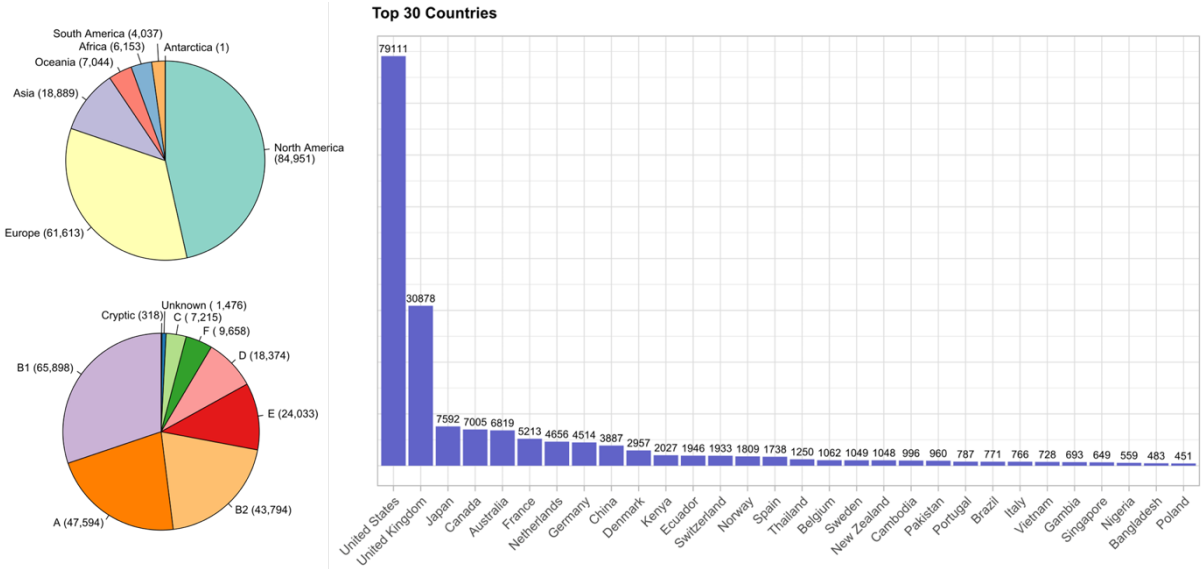
PSIPRED protein structure predictions, and associated membrane helix prediction using MEMSAT-SVM. CDS42, CDS43, CDS58, CDS61, CDS146, CDS201 and CDS57 were all too short for analysis (minimum size is 30 residues).



Supplementary Figure 12. Predicted protein structures
Protein structures predicted using AlphaFold. All structures are shown with the N-terminus on the left and coloured by pLDDT using Chimera.



Supplementary Figure 13. Multisequence alignment plots from the AlphaFold predictions



Supplementary Figure 14. Distribution of reference genomes by Continent, Country and Phylogroup.

732 **References**

733 1. Ruiz-Orera, J. & Albà, M. M. Translation of Small Open Reading Frames: Roles in
734 Regulation and Evolutionary Innovation. *Trends Genet.* **35**, 186–198 (2019).
735 2. Makarewich, C. A. & Olson, E. N. Mining for Micropeptides. *Trends Cell Biol.* **27**, 685–696
736 (2017).
737 3. Hellens, R. P., Brown, C. M., Chisnall, M. A. W., Waterhouse, P. M. & Macknight, R. C.
738 The Emerging World of Small ORFs. *Trends Plant Sci.* **21**, 317–328 (2016).
739 4. Storz, G., Wolf, Y. I. & Ramamurthi, K. S. Small Proteins Can No Longer Be Ignored. *Annu.*
740 *Rev. Biochem.* **83**, 753–777 (2014).
741 5. Islam, M. S., Shaw, R. K., Frankel, G., Pallen, M. J. & Busby, S. J. W. Translation of a
742 minigene in the 5' leader sequence of the enterohaemorrhagic Escherichia coli LEE1
743 transcription unit affects expression of the neighbouring downstream gene. *Biochem. J.*
744 **441**, 247–53 (2012).
745 6. Dar, D. *et al.* Term-seq reveals abundant ribo-regulation of antibiotics resistance in
746 bacteria. *Science* **352**, aad9822–aad9822 (2016).
747 7. Kędzierska, B. & Potrykus, K. Minigene as a Novel Regulatory Element in Toxin-Antitoxin
748 Systems. *Int. J. Mol. Sci.* **22**, 13389 (2021).
749 8. Duval, M. & Cossart, P. Small bacterial and phagic proteins: an updated view on a rapidly
750 moving field. *Curr. Opin. Microbiol.* **39**, 81–88 (2017).
751 9. Hemm, M. R., Weaver, J. & Storz, G. Escherichia coli small proteome. *EcoSal Plus* **9**,
752 (2020).
753 10. Gray, T., Storz, G. & Papenfort, K. Small Proteins; Big Questions. *J. Bacteriol.* **204**, (2022).

754 11. Impens, F. *et al.* N-terminomics identifies Prli42 as a membrane miniprotein conserved
755 in Firmicutes and critical for stressosome activation in Listeria monocytogenes. *Nat.*
756 *Microbiol.* **2**, 17005 (2017).
757 12. Unoson, C. & Wagner, E. G. H. A small SOS-induced toxin is targeted against the inner
758 membrane in *Escherichia coli*. *Mol. Microbiol.* **70**, 258–270 (2008).
759 13. Su, W.-L. *et al.* TisB Protein Protects Escherichia coli Cells Suffering Massive DNA
760 Damage from Environmental Toxic Compounds. *mBio* **13**, (2022).
761 14. Karimova, G., Davi, M. & Ladant, D. The β-Lactam Resistance Protein Blr, a Small
762 Membrane Polypeptide, Is a Component of the Escherichia coli Cell Division Machinery.
763 *J. Bacteriol.* **194**, 5576–5588 (2012).
764 15. Wong, R. S. Y., McMurry, L. M. & Levy, S. B. 'Intergenic' *blr* gene in *Escherichia coli*
765 encodes a 41-residue membrane protein affecting intrinsic susceptibility to certain
766 inhibitors of peptidoglycan synthesis. *Mol. Microbiol.* **37**, 364–370 (2000).
767 16. Handler, A. A., Lim, J. E. & Losick, R. Peptide inhibitor of cytokinesis during sporulation in
768 *Bacillus subtilis*. *Mol. Microbiol.* **68**, 588–599 (2008).
769 17. Modell, J. W., Hopkins, A. C. & Laub, M. T. A DNA damage checkpoint in *Caulobacter*
770 *crescentus* inhibits cell division through a direct interaction with FtsW. *Genes Dev.* **25**,
771 1328–1343 (2011).
772 18. Lloyd, C. R., Park, S., Fei, J. & Vanderpool, C. K. The Small Protein SgrT Controls Transport
773 Activity of the Glucose-Specific Phosphotransferase System. *J. Bacteriol.* **199**, (2017).
774 19. Martin, J. E., Waters, L. S., Storz, G. & Imlay, J. A. The Escherichia coli Small Protein MntS
775 and Exporter MntP Optimize the Intracellular Concentration of Manganese. *PLOS Genet.*
776 **11**, e1004977 (2015).

777 20. Kohlstaedt, M. *et al.* Identification and Characterization of the Novel Subunit CcoM in
778 the *cbb* $_3$ -Cytochrome *c* Oxidase from Pseudomonas stutzeri ZoBell. *mBio* **7**, (2016).
779 21. VanOrsdel, C. E. *et al.* The Escherichia coli CydX protein is a member of the CydAB
780 cytochrome bd oxidase complex and is required for cytochrome bd oxidase activity. *J.*
781 *Bacteriol.* **195**, 3640–50 (2013).
782 22. Lippa, A. M. & Goulian, M. Feedback Inhibition in the PhoQ/PhoP Signaling System by a
783 Membrane Peptide. *PLoS Genet.* **5**, e1000788 (2009).
784 23. Rowland, S. L. *et al.* Structure and mechanism of action of Sda, an inhibitor of the
785 histidine kinases that regulate initiation of sporulation in Bacillus subtilis. *Mol. Cell* **13**,
786 689–701 (2004).
787 24. Hobbs, E. C., Yin, X., Paul, B. J., Astarita, J. L. & Storz, G. Conserved small protein
788 associates with the multidrug efflux pump AcrB and differentially affects antibiotic
789 resistance. *Proc. Natl. Acad. Sci.* **109**, 16696–16701 (2012).
790 25. Laczkovich, I. *et al.* Discovery of Unannotated Small Open Reading Frames in
791 Streptococcus pneumoniae D39 Involved in Quorum Sensing and Virulence Using
792 Ribosome Profiling. *mBio* **13**, e01247-22 (2022).
793 26. Gannoun-Zaki, L., Alibaud, L., Carrère-Kremer, S., Kremer, L. & Blanc-Potard, A.-B.
794 Overexpression of the KdpF Membrane Peptide in Mycobacterium bovis BCG Results in
795 Reduced Intramacrophage Growth and Altered Cording Morphology. *PLoS ONE* **8**,
796 e60379 (2013).
797 27. Miravet-Verde, S. *et al.* Unraveling the hidden universe of small proteins in bacterial
798 genomes. *Mol. Syst. Biol.* **15**, (2019).
799 28. Crappé, J. *et al.* Combining in silico prediction and ribosome profiling in a genome-wide
800 search for novel putatively coding sORFs. *BMC Genomics* **14**, 648 (2013).

801 29. Meydan, S. *et al.* Retapamulin-assisted ribosome profiling reveals the alternative
802 bacterial proteome. *Mol. Cell* **74**, 481 (2019).
803 30. Smith, R. A. & Parkinson, J. S. Overlapping genes at the cheA locus of Escherichia coli.
804 *Proc. Natl. Acad. Sci. U. S. A.* **77**, 5370–5374 (1980).
805 31. Nyengaard, N. R., Mortensen, K. K., Lassen, S. F., Hershey, J. W. & Sperling-Petersen, H.
806 U. Tandem translation of E. coli initiation factor IF2 beta: purification and
807 characterization in vitro of two active forms. *Biochem. Biophys. Res. Commun.* **181**,
808 1572–1579 (1991).
809 32. Miller, M. J. & Wahba, A. J. Chain initiation factor 2. Purification and properties of two
810 species from Escherichia coli MRE 600. *J. Biol. Chem.* **248**, 1084–1090 (1973).
811 33. DiMaio, D. Viral miniproteins. *Annu. Rev. Microbiol.* **68**, 21–43 (2014).
812 34. Vila, J., Vargas, M., Henderson, I. R., Gascón, J. & Nataro, J. P. Enteroaggregative
813 Escherichia coli Virulence Factors in Traveler's Diarrhea Strains. *J. Infect. Dis.* **182**, 1780–
814 1783 (2000).
815 35. D'Souza, C., Nakano, M. M. & Zuber, P. Identification of comS, a gene of the srfA operon
816 that regulates the establishment of genetic competence in Bacillus subtilis. *Proc. Natl.*
817 *Acad. Sci.* **91**, 9397–9401 (1994).
818 36. Kreitmeier, M. *et al.* Spotlight on alternative frame coding: Two long overlapping genes
819 in Pseudomonas aeruginosa are translated and under purifying selection. *iScience* **25**,
820 103844 (2022).
821 37. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide
822 Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling.
823 *Science* **324**, 218–223 (2009).

824 38. Weaver, J., Mohammad, F., Buskirk, A. R. & Storz, G. Identifying Small Proteins by
825 Ribosome Profiling with Stalled Initiation Complexes. *mBio* **10**, (2019).
826 39. Mohammad, F., Green, R. & Buskirk, A. R. A systematically-revised ribosome profiling
827 method for bacteria reveals pauses at single-codon resolution. *eLife* **8**, (2019).
828 40. Vazquez-Laslop, N., Sharma, C. M., Mankin, A. & Buskirk, A. R. Identifying Small Open
829 Reading Frames in Prokaryotes with Ribosome Profiling. *J. Bacteriol.* **204**, e00294-21
830 (2022).
831 41. Stringer, A., Smith, C., Mangano, K. & Wade, J. T. Identification of Novel Translated Small
832 Open Reading Frames in *Escherichia coli* Using Complementary Ribosome Profiling
833 Approaches. *J. Bacteriol.* **204**, (2022).
834 42. Ahrens, C. H., Wade, J. T., Champion, M. M. & Langer, J. D. A Practical Guide to Small
835 Protein Discovery and Characterization Using Mass Spectrometry. *J. Bacteriol.* **204**,
836 (2022).
837 43. Hemm, M. R. *et al.* Small Stress Response Proteins in Escherichia coli: Proteins Missed by
838 Classical Proteomic Studies. *J. Bacteriol.* **192**, 46–58 (2010).
839 44. Manoil, C. & Beckwith, J. TnphoA: a transposon probe for protein export signals. *Proc.*
840 *Natl. Acad. Sci. U. S. A.* **82**, 8129–33 (1985).
841 45. Goodall, E. C. A. *et al.* The essential genome of Escherichia coli K-12. *mBio* **9**, (2018).
842 46. Yakhnin, H., Aichele, R., Ades, S. E., Romeo, T. & Babitzke, P. Circuitry Linking the Global
843 Csr- and σ^E -Dependent Cell Envelope Stress Response Systems. *J. Bacteriol.* **199**, (2017).
844 47. Schmidt, A. *et al.* The quantitative and condition-dependent *Escherichia coli* proteome.
845 *Nat. Biotechnol.* **34**, 104–110 (2016).
846 48. Fuchs, S. & Engelmann, S. Small proteins in bacteria - Big challenges in prediction and
847 identification. *Proteomics* **23**, e2200421 (2023).

848 49. Fijalkowski, I., Willems, P., Jonckheere, V., Simoens, L. & Van Damme, P. Hidden in plain
849 sight: challenges in proteomics detection of small ORF-encoded polypeptides. *microLife*
850 **3**, uqac005 (2022).
851 50. Hecht, A. *et al.* Measurements of translation initiation from all 64 codons in E. coli.
852 *Nucleic Acids Res.* **45**, 3615–3626 (2017).
853 51. Baggett, N. E., Zhang, Y. & Gross, C. A. Global analysis of translation termination in E.
854 coli. *PLOS Genet.* **13**, e1006676 (2017).
855 52. VanOrsdel, C. E. *et al.* Identifying New Small Proteins in *Escherichia coli*. *PROTEOMICS*
856 **18**, (2018).
857 53. Hemm, M. R., Paul, B. J., Schneider, T. D., Storz, G. & Rudd, K. E. Small membrane
858 proteins found by comparative genomics and ribosome binding site models. *Mol.*
859 *Microbiol.* **70**, 1487–501 (2008).
860 54. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein
861 language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).
862 55. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**,
863 679–682 (2022).
864 56. Li, E. H. *et al.* Blind assessment of monomeric AlphaFold2 protein structure models with
865 experimental NMR data. *J. Magn. Reson. San Diego Calif 1997* **352**, 107481 (2023).
866 57. Steinberg, R. & Koch, H.-G. The largely unexplored biology of small proteins in pro- and
867 eukaryotes. *FEBS J.* **288**, 7002–7024 (2021).
868 58. Garai, P. & Blanc-Potard, A. Uncovering small membrane proteins in pathogenic
869 bacteria: Regulatory functions and therapeutic potential. *Mol. Microbiol.* **114**, 710–720
870 (2020).

871 59. Burton, A. T., Zeinert, R. & Storz, G. Large Roles of Small Proteins. *Annu. Rev. Microbiol.*
872 (2024) doi:10.1146/annurev-micro-112723-083001.
873 60. Wang, H. *et al.* Increasing intracellular magnesium levels with the 31-amino acid MgtS
874 protein. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5689–5694 (2017).
875 61. Nord, S., Bylund, G. O., Lövgren, J. M. & Wikström, P. M. The RimP protein is important
876 for maturation of the 30S ribosomal subunit. *J. Mol. Biol.* **386**, 742–753 (2009).
877 62. Higashi, K. *et al.* Identification of a spermidine excretion protein complex (MdtJI) in
878 Escherichia coli. *J. Bacteriol.* **190**, 872–878 (2008).
879 63. Sperandeo, P. *et al.* Characterization of lptA and lptB, two essential genes implicated in
880 lipopolysaccharide transport to the outer membrane of Escherichia coli. *J. Bacteriol.*
881 **189**, 244–253 (2007).
882 64. Lee, C.-R., Cho, S.-H., Yoon, M.-J., Peterkofsky, A. & Seok, Y.-J. Escherichia coli enzyme
883 IIANtr regulates the K+ transporter TrkA. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 4124–4129
884 (2007).
885 65. Kalamorz, F., Reichenbach, B., März, W., Rak, B. & Görke, B. Feedback control of
886 glucosamine-6-phosphate synthase GlmS expression depends on the small RNA GlmZ
887 and involves the novel protein YhbJ in Escherichia coli. *Mol. Microbiol.* **65**, 1518–1533
888 (2007).
889 66. Zallot, R., Yuan, Y. & de Crécy-Lagard, V. The Escherichia coli COG1738 Member YhhQ Is
890 Involved in 7-Cyanodeazaguanine (preQ$_0$) Transport. *Biomolecules* **7**, (2017).
891 67. Neuhaus, K. *et al.* Translatomics combined with transcriptomics and proteomics reveals
892 novel functional, recently evolved orphan genes in Escherichia coli O157:H7 (EHEC).
893 *BMC Genomics* **17**, 133 (2016).

894 68. Miravet-Verde, S. *et al.* ProTInSeq: transposon insertion tracking by ultra-deep DNA
895 sequencing to identify translated large and small ORFs. *Nat. Commun.* **15**, 2091 (2024).
896 69. Ardern, Z., Neuhaus, K. & Scherer, S. Are Antisense Proteins in Prokaryotes Functional?
897 *Front. Mol. Biosci.* **7**, (2020).
898 70. Meydan, S., Vázquez-Laslop, N. & Mankin, A. S. Genes within Genes in Bacterial
899 Genomes. *Microbiol. Spectr.* **6**, 10.1128/microbiolspec.rwr-0020–2018 (2018).
900 71. Lawaetz, A. C., Cowley, L. A. & Denham, E. L. Genome-wide annotation of transcript
901 boundaries using bacterial Rend-seq datasets. *Microb. Genomics* **10**, 001239 (2024).
902 72. Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**,
903 6614–6624 (2016).
904 73. Angiuoli, S. V. *et al.* Toward an Online Repository of Standard Operating Procedures
905 (SOPs) for (Meta)genomic Annotation. *OMICS J. Integr. Biol.* **12**, 137–141 (2008).
906 74. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in
907 *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6640–5
908 (2000).
909 75. Goodall, E. C. A. *et al.* A multiomic approach to defining the essential genome of the
910 globally important pathogen Corynebacterium diphtheriae. *PLOS Genet.* **19**, e1010737
911 (2023).
912 76. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinforma. Oxf.*
913 *Engl.* **16**, 944–5 (2000).
914 77. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
915 features. *Bioinformatics* **26**, 841–842 (2010).
916 78. Zeghouf, M. *et al.* Sequential Peptide Affinity (SPA) System for the Identification of
917 Mammalian and Bacterial Protein Complexes. *J. Proteome Res.* **3**, 463–468 (2004).

918 79. Venturini, E. *et al.* A global data-driven census of *Salmonella* small proteins and their
919 potential functions in bacterial virulence. *microLife* **1**, (2020).
920 80. Hilgarth, R. S. & Lanigan, T. M. Optimization of overlap extension PCR for efficient
921 transgene construction. *MethodsX* **7**, 100759 (2020).
922 81. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout
923 mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
924 82. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on.
925 *Nucleic Acids Res.* **47**, (2019).
926 83. Nugent, T. & Jones, D. T. Transmembrane protein topology prediction using support
927 vector machines. *BMC Bioinformatics* **10**, 159 (2009).
928 84. Zhou, Z. *et al.* The EnteroBase user's guide, with case studies on Salmonella
929 transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity.
930 *Genome Res.* **30**, 138–152 (2020).
931