https://doi.org/10.1093/nar/gkac302

"Pseudo-pseudogenes" in bacterial genomes: Proteogenomics reveals a wide but low protein expression of pseudogenes in *Salmonella enterica*

Ye Feng[1,2,*], Zeyu Wang[1,2], Kun-Yi Chien[3], Hsiu-Ling Chen[4], Yi-Hua Liang[4], Xiaoting Hua[1] and Cheng-Hsun Chiu[3,4,5,*]

[1]Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, People's Republic of China, [2]Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, People's Republic of China, [3]Graduate Institute of Biomedical Sciences, Chang Gung University College of Medicine, Taoyuan, Republic of China, [4]Molecular Infectious Disease Research Center, Chang Gung Memorial Hospital, Taoyuan, Republic of China and [5]Division of Pediatric Infectious Diseases, Department of Pediatrics, Chang Gung Memorial Hospital, Chang Gung University College of Medicine, Taoyuan, Republic of China

Received December 07, 2021; Revised April 11, 2022; Editorial Decision April 12, 2022; Accepted April 14, 2022

ABSTRACT

Pseudogenes (genes disrupted by frameshift or in-frame stop codons) are ubiquitously present in the bacterial genome and considered as nonfunctional fossil. Here, we used RNA-seq and mass-spectrometry technologies to measure the transcriptomes and proteomes of *Salmonella enterica* serovars Paratyphi A and Typhi. All pseudogenes' mRNA sequences remained disrupted, and were present at comparable levels to their intact homologs. At the protein level, however, 101 out of 161 pseudogenes suggested successful translation, with their low expression regardless of growth conditions, genetic background and pseudogenization causes. The majority of frameshifting detected was compensatory for -1 frameshift mutations. Readthrough of in-frame stop codons primarily involved UAG; and cytosine was the most frequent base adjacent to the codon. Using a fluorescence reporter system, fifteen pseudogenes were confirmed to express successfully *in vivo* in *Escherichia coli*. Expression of the intact copy of the fifteen pseudogenes in *S.* Typhi affected bacterial pathogenesis as revealed in human macrophage and epithelial cell infection models. The above findings suggest the need to revisit the nonstandard translation mechanism as well as the biological role of pseudogenes in the bacterial genome.

INTRODUCTION

Pseudogenes (ψ) are DNA sequences that resemble genes but have been mutated into defective forms over the course of evolution. Within bacterial and viral genomes, pseudogenes refer mainly to those disrupted by frameshift and nonsense mutation (1–3). In contrast, within eukaryotic genomes, such pseudogenes are categorized as 'unitary' (4), and more pseudogenization events arise from retrotransposition of mRNA sequences into genomes in the case of 'processed' pseudogenes, and genomic duplication in the case of 'duplicated' pseudogenes (5). Pseudogenes exist across almost all forms of life and vary in number between organisms. Although the majority of bacterial genomes are compact and have a high gene density (2,3,6), there are many intracellular bacteria, such as *Sodalis glossinidius* (7) and *Mycobacterium leprae* (8) whose coding capacity is greatly reduced compared to their free-living ancestors.

Pseudogenization in bacteria is primarily regarded as a process in which certain genes become disposable to the fitness of the organism due to a shift in niche and hence loss of selective pressure (2,9). This have been well illustrated by genomic comparison between different serovars of *Salmonella enterica*. As a gram-negative facultative anaerobe, while most of its 1,600 serovars are host-generalists, a few serovars are host-specialists, such as *S.* Typhi and *S.* Paratyphi A that are restricted to infecting human, and *S.* Gallinarum to infecting fowl (10–12). As pseudogenes in the host-specialist serovars greatly outnumber that in the host-generalist serovars *S.* Typhimurium and *S.* Enteritidis (13–18), pseudogenization has been considered a hallmark of the host-specialists. On the other hand, as argued by the 'less is more' hypothesis, pseudogenization of certain genes could be advantageous to the mutant and con-

[*]To whom correspondence should be addressed. Tel: +86 571 88981576; Fax: +86 571 88981576; Email: pandafengye@zju.edu.cn
Correspondence may also be addressed to Cheng-Hsun Chiu. Tel: +886 3 3281200; Fax: +886 3 3288957; Email: chchiu@adm.cgmh.org.tw

sequently sweep through the population to fixation. For example, pseudogenization of the *Salmonella* pathogenicity island 1 (SPI-1) effector gene *sopE2* and *sseJ* facilitates invasion of *S.* Typhi into epithelial cells and therefore contributes to adaptation of *S.* Typhi in the systemic infection of humans (19,20).

An increasing number of studies have provided evidence supporting a biological function of pseudogenes. In addition to the regulatory roles mediated by the mRNA of pseudogenes, a few pseudogenes can be translated to synthesize a full-length product (21,22). This so-called 'recoding' process and the resulting 'pseudo-pseudogenes' thereby confute the originally pejorative inference of the 'pseudogene' label (23). Translational recoding is achieved mainly by codon redefinition and frameshifting. Codon redefinition results either from an erroneously charged tRNA or an anticodon-codon mismatch on the ribosome (24). The amino acids specified for the in-frame stop codon are often either tryptophan or glutamine depending on whether UGA or UAG are being readthrough (24). Frameshifting refers mostly to repositioning of the ribosome by one nucleotide backward (−1 frameshifting) or forward (+1 frameshifting) so that translation can continue in the new reading frame; but − 2 frameshifting has also been reported (25). For frameshift mutations within mononucleotide repeats, transcriptional slippage can incorporate or remove one or more bases in mRNA and therefore restore the open reading frame (25).

Despite being outside the mainstream of genetic mechanisms, growing evidence suggests recoding generates phenotypic heterogeneity and hence helps the bacterial population quickly adjust to environmental changes. For example, recoding in the β-subunit of RNA polymerase (RpoB) in *Escherichia coli* and *Mycobacterium tuberculosis* increases the resistance to rifampicin (26,27). For genes like *dnaX*, *prfB* and *shdA*, one gene yields two protein isoforms with different functions, with one whose synthesis utilizes recoding at the in-frame 'disruptive' site and the other not (28,29). Bacterial phenotype can be affected by the ratio of the two isoforms, which is further modulated by recoding. Under such circumstances, these genes should not be termed pseudogenes by strict definition; but for the sake of simplicity, we do not make a clear distinction in this study.

For bacteria, recoding is relatively abundant among *prfB*, *dnaX*, *rpoB*, and mobile genes, namely those encoding integrase/transposase (30–34). Leaving aside the mobile genes which are probably not of eubacterial origin, it is unclear whether recoding is rare for bacteria. The answer to this question is of particular importance since pseudogenization sheds light on bacterial evolution concerning how bacteria have adapted to novel environments. More importantly, it will also enlarge the knowledge on which non-standard transcription/translation mechanisms are recruited to overcome disruptive codons and restore the original functions of pseudogenes. A few *in silico* studies have indicated the presence of eubacterial recoding (31,35), but the wet-lab evidence is scarce. In the recent decade, proteomics has been applied to identification of translated pseudogenes in bacteria, including *M. tuberculosis*, *S. glossinidius*, *Shewanella oneidensis*, and *Yersinia* strains (36–39). While peptides derived from dozens of pseudogenes were identified, few of these studies inspected the peptide location in relation to the disruptive site to rule out the possibility of translational reinitiation with an alternative start codon, or validated the pseudogene expression by other methods. Thus these findings fall short of proving the successful readthrough of the disruptive sites, let alone investigating the recoding mechanism.

Here, we utilized RNA sequencing and proteomic analysis to fully describe the transcriptional and translational landscape of three *Salmonella* serovars under two different growth conditions, and to identify the potential recoding events. We also applied a fluorescence reporter system to double validate the expression of pseudogenes, and found them significantly lower than the intact homologs. This wide but low protein expression of pseudogenes prompts us to reconsider the concept of pseudogene and its functional and evolutionary role in bacterial genomes.

MATERIALS AND METHODS

In silico prediction of pseudogenes

The accession number of the genome sequences representing *S.* Typhimurium, *S.* Typhi and *S.* Paratyphi A are NC_016856 (strain 14028S), NC_003198 (strain CT18), NC_006511 (strain ATCC9150). Orthologous relationship between strains were constructed using the Roary program v3.11.2 (40). A manual determination on whether the genes were intact was performed, with the genes of the *S.* Typhimurium strain 14028S were used as a reference. Genes that remained intact in *S.* Typhimurium but contained frameshifts or in-frame stop codons in *S.* Typhi or *S.* Paratyphi A were investigated in this study. The exact disruptive site was identified using GeneWise software v2.4.1 (41). Genes with multiple copies in the chromosome or with imperfect synteny were excluded from further analysis.

Bacterial strains and culture

S. Typhimurium strain 14028S, *S.* Typhi strain CT18 and *S.* Paratyphi A strain ATCC9150 were enrolled in this study. Single colonies of these strains were cultured in LB medium overnight, and then was transferred into fresh LB medium and magnesium minimal medium (MgM medium, pH 7.4) (42), respectively, with a dilution ratio of 1:100. Till OD_{600} reached 0.8–0.9 the bacteria were harvested for proteomic and transcriptomic analyses.

Liquid chromatography with tandem mass spectrometry (LC-MS/MS)

Bacterial proteins were extracted with 0.1% SDS solution by sonication. Protein concentration was determined by the BCA method (Thermo Fisher, US). Forty micrograms of protein extracts were reduced and alkylated with dithiothreitol and iodoacetamide, respectively, and digested with 1 μg of modified trypsin (Promega, US) at 37°C overnight. The peptides were then labeled using TMT 10 plex Mass Tag Labeling Kits (Thermo Fisher, US) according to the manufacturer's instructions.

The comprehensive 2D-SCX-RP-LC system (Ultimate 3000, Germany) was equipped with one gradient pump for strong cation exchange (SCX), one other gradient pump for reversed phase (RP), one isocratic pump for online dilution, one 10-port valve with two RP-trapping columns for alternating trapping, one 6-port valve for controlling the trapping column washing, and a manual injector for sample loading. Such a combination allowed us to introduce an organic solvent (acetonitrile) in the first dimensional SCX separation without affecting the second dimensional RP separations by using an online dilution design. Briefly, samples dissolved in 50% acetonitrile containing 0.1% formic acid were loaded onto the SCX column through a manual injector. The flow rate of the first dimensional separation was operated at a flow rate of 1μL/min on a home-packed SCX column (0.5 × 150 mm, pack with Luna-SCX particles from Phenomenex, US). The peptides were eluted using a continuous ammonium chloride concentration gradient in the presence of 0.1% formic acid and 30% ACN. The salt gradient was segmented in 17 steps, 90 min for each, and matched with the second dimensional reverse phase separations. The isocratic loading pump delivering 50 μL/min of solvent A (0.1% formic acid in water) was used for diluting the effluent of SCX column through a T-union and mixing tubing before it reached the trapping column. In the meantime, the other RP-trapping column, installed on the same 10-port valve, was connected with the RP-separating column and was being analyzed by a mass spectrometer. Six minutes before each salt gradient step being completed, the binary pump for the SCX separation stopped, and the six-port valve switched to allow the loading pump to wash away the residual salt solution in the flow path of the RP-trapping column. After the trapping column switched to the RP analytical column, the bound peptides were eluted with a complete acetonitrile gradient (elution, regeneration, and then re-equilibration) in the presence of 0.05% formic acid over 84 min.

The effluent of the online 2D LC was analyzed by a LTQ-Orbitrap hybrid mass spectrometer (Thermo Electron, Germany). The mass spectrometer was operated in the information-dependent acquisition (IDA) mode. Survey full scan MS spectra (from m/z300 to 2000) are acquired in the Orbitrap at a resolution of 60,000 with lock mass function enabled. Ten most intense ions in each MS spectrum are selected for isolation and fragmentation in the linear ion trap (MS/MS). Each precursor ion was allowed to be analyzed twice and then excluded in the subsequent 1 min. The MS/MS isolation width was set to 2 Da, and the maximum precursor accumulation time for MS/MS was set to 150 ms.

Raw MS files from the LTQ-Orbitrap were analyzed by Mascot v2.2.2 (Matrix Science, US) and MaxQuant v1.0.13.13 (Matrix Science, US). MS/MS spectra were searched with Mascot using an in-house sequence database. Briefly, the protein-coding sequences of the three *Salmonella* genomes were extracted from Genbank files. Their non-redundant protein sequences, based on the orthologous relationship built by Roary as described above, were used as database for proteome search. For pseudogenes, twenty possible amino acids were individually replaced at the disruptive site of the protein sequences of the intact homologs, and all of the allele sequences were put into the database. Peptides translated from the annotated pseudogenes following the standard genetic code were also put into the database. The parameters setting for the Mascot searches are as follows: cysteine carbamidomethylation was selected as a fixed modification, whereas protein N-terminal acetylation and methionine oxidation were selected as variable modifications, and a maximum of two missed cleavages were allowed. Parent mass and fragment ions were searched with mass deviation of 5 ppm and 0.5 Da, respectively. The search results were further processed by MaxQuant with the following parameter settings: peptides of minimum six amino acids were allowed, false discovery rate was set to 0.01, and a posterior error probability (PEP) for each MS/MS spectrum below or equal to 0.1 was required. The detected peptides mapped against the pseudogenes were further searched against the database with NCBI tblastn program to make sure to have a unique hit in the database.

RNA-seq

Total RNA was isolated using MicroRNA easy Kit (Qiagen, Germany) following the manufacturer's instructions. The ribosomal RNA was removed with Ribo-Zero rRNA Removal Kit (Bacteria) (Illumina, US). Sequencing library was constructed using VAHTSTM Stranded mRNA-seq Library Prep Kit for Illumina (Vazyme Biotech, China). The library was then sequenced on an Illumina HiSeq4000 platform, generating 100 bp paired-end reads.

Bowtie2 v2.3.5 (43) was used to map the raw reads against the reference genomes (see accession number above). The normalized expression abundance of each gene was estimated by RSEM v1.3.1 (44). The Integrative Genomics Viewer v2.3 (45) was used to view the alignment and check if the transcribed sequences were identical to genome sequence and if any mutation events occurred at the transcription stage. The transcription start sites were predicted using Rockhopper software (46); the transcriptome data for all strains used the genome of 14028S (accession no. NC_016856) as the reference. The prediction results were listed in the Supplementary Table S1.

Identification of motifs around the disruptive site

Nucleotide sequences (45 bp upstream and downstream of the disruptive site, respectively) were used as input to the online MEME v4.12.0 for identification of the conserved sequence structure around the mutation site (47); the width of motif was set between 5–15 bp; each motif should have at least four sites. Selenocysteine insertion sequence (SECIS) elements were predicted by the bSECISearch program (48). Simple tandem repeat sequences were identified using SSRIT software (49); maximum length of motif was set as three, and minimum number of repeats was five.

Expression of pseudogenes in *E. coli* and flow cytometry analysis

The pseudogenes and their intact homologs were amplified by PCR using the chromosomal DNA of *S.* Typhi strain CT18 and *S.* Typhimurium strain 14028S as template; the primer sets were listed in Supplementary Table S2. The purified PCR products were cloned into the pEL-polB-sYFP2 plasmid vector using Hieff Clone® Plus Multi One

Step Cloning Kit (Yesen, China) (50). A linker sequence (5′-GGAGGAGGAGGAAGC-3′) was used to link the test gene (the termination codon is removed) with the *syfp2* gene to form a fusion gene.

The recombinant plasmid was then transformed into *E. coli* DH5α. The resulting transformant was confirmed by PCR and Sanger sequencing. Next, to check if the disruptive mutation was still present at the RNA level, the transformant was cultured to exponential phase. The RNA was extracted using TRLzol, treated with DNase (Thermo Fisher, US), and purified with RNeasy kit (Qiagen, Germany). The cDNA was synthesized using iScript cDNA Synthesis Kit (Bio-Rad, US) and sequenced by Sanger sequencing.

Flow cytometry was performed to test the *E. coli* population tagged with the yellow fluorescent SYFP2 protein. Overnight cultures in LB broth were diluted 1:100 in phosphate buffer saline (PBS) and then were analyzed with BD FACS Calibur (Becton, Dickinson and Company, US). *E. coli* strains DH5α and XH141 were used as negative control and positive control respectively (50).

To further confirm the expression of pseudogenes, the PCR products of the pseudogenes and their intact homologs were cloned into vector pET29b (kanamycin-resistant) upstream of the His-tag site. The plasmid was introduced into *E.coli* BL21 (DE3) through electrotransformation. The transformant were cultured overnight, refreshed 1:100 into LB broth, and cultured till OD_{600} approached to 0.6. Then isopropyl β-d-1-thiogalactopyranoside (IPTG) was added to induce protein production. The induced bacteria were lysed by using B-PER bacterial protein extraction reagent (Thermo Fisher, US) and centrifuged. The supernatants were added to the prepacked His GraviTrap™ TALON® columns (Merck KGaA, Germany), and the target proteins were collected with the manufacturer's instruction. The protein elution was separated by SDS-PAGE and the protein bands at appropriate size were cut off for the protein identification, which was conducted with Microflex matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS, Bruker Daltonics, US).

Bile tolerance assay

Bacterial culture was grown overnight and diluted to 1:100 in MH broth with 5% pig bile salt (SolarBio, China), and then the aliquots were placed into a flat-bottom plate. The plate was incubated at 37°C with agitation. The OD_{600} value of each culture was continuously determined for 20 h using Bioscreen C MBR instrument (Oy Growth Curves Ab Ltd., Finland). The growth rate was estimated based on OD_{600} curves using an R script as previously described (51), and then was compared using one-way analysis of variance followed by Tukey's multiple comparisons tests.

Bacterial invasion assay and intracellular survival assay

The plasmids expressing the intact copy of the query pseudogenes, which were constructed as described above, were electrotransformed into *S.* Typhi strain CMCC50097. For the bacterial invasion assay, the human colon cancer cell line Caco-2 was maintained in Dulbecco's modified eagle medium (DMEM, Gibco, US) supplemented with 1% nonessential amino acids (SolarBio, China) and 20% fetal bovine serum (Cellmax, China) at 5% CO2 and 37°C. The Caco-2 cells were seeded onto 24-well plates at a density of 2×10^5 cells/well. Then the mid-log-phase bacteria were added to the cell monolayer at an multiplicity of infection (MOI) of 50 : 1, and were centrifuged at 800 *g* for 5 min. After 2 h of incubation, the bacteria not invading in the cells were removed by washing with PBS. Subsequently, the cells were further incubated in DMEM supplemented with 100 mg/mL gentamicin for 30 min to kill extracellular bacteria and again in DMEM supplemented with 5 mg/mL gentamicin for 2 h. Finally, cells were washed with PBS, lysed with 0.1% Triton-100X; the lysate was diluted and plated on LB agar to determine the number of survived bacteria. The relative invasion ability was defined as the number of bacteria recovered 24 h for each mutant strain divided by that for the wild strain.

The human monocyte cell line THP-1 was used for the intracellular survival assay. Before infection, cells were cultured in RPMI 1640 medium (Gibco, US) supplemented with 2-mercaptoethanol and then differentiated by addition of phorbol 12-myristate 13-acetate (PMA) for 48 h. The protocol of the infection experiment was the same as the above invasion assay, except that the MOI was 100:1 and the initial incubation was 1 h.

RESULTS

Pseudogenes remain disrupted at the RNA level

The *in silico* prediction of pseudogene was performed among the positional orthologous genes in *Salmonella* serovars. Using the sequence of the host-generalist serovar *S.* Typhimurium as the control, 94 and 73 pseudogenes were identified from the host-specialist serovars *S.* Paratyphi A and *S.* Typhi via frameshift or nonsense mutation. Six pseudogenes were disrupted in both of the two human-restricted serovars, suggesting their convergent adaptation to the host-restricted niche.

The SPI-1- and SPI-2-inducing media, which mimicked two representative *Salmonella* infection environments, i.e. intestinal inflammation and sustained intracellular colonization, were used to culture the *Salmonella* strains (52,53). RNA-seq was performed upon the harvested bacteria. All but one of the 161 pseudogenes had their transcriptomic reads mapped by at least ten RNA-seq reads, and the sequences of these reads at the disruptive sites were identical to their DNA templates. These results suggested that pseudogenes in *Salmonella* could be transcribed and that the open reading frames remained disrupted at the RNA level.

Pseudogenes can express full-length proteins

The proteomes of bacteria harvested under the two culturing conditions were investigated by LC-MS/MS technology. Out of the 4200–4400 annotated protein-encoding genes, 3298, 3209 and 3261 proteins were identified for the *S.* Typhimurium, *S.* Paratyphi A and *S.* Typhi proteomes, respectively. Of the 161 annotated pseudogenes, 101 had detectable peptides (Supplementary Table S3). BLAST

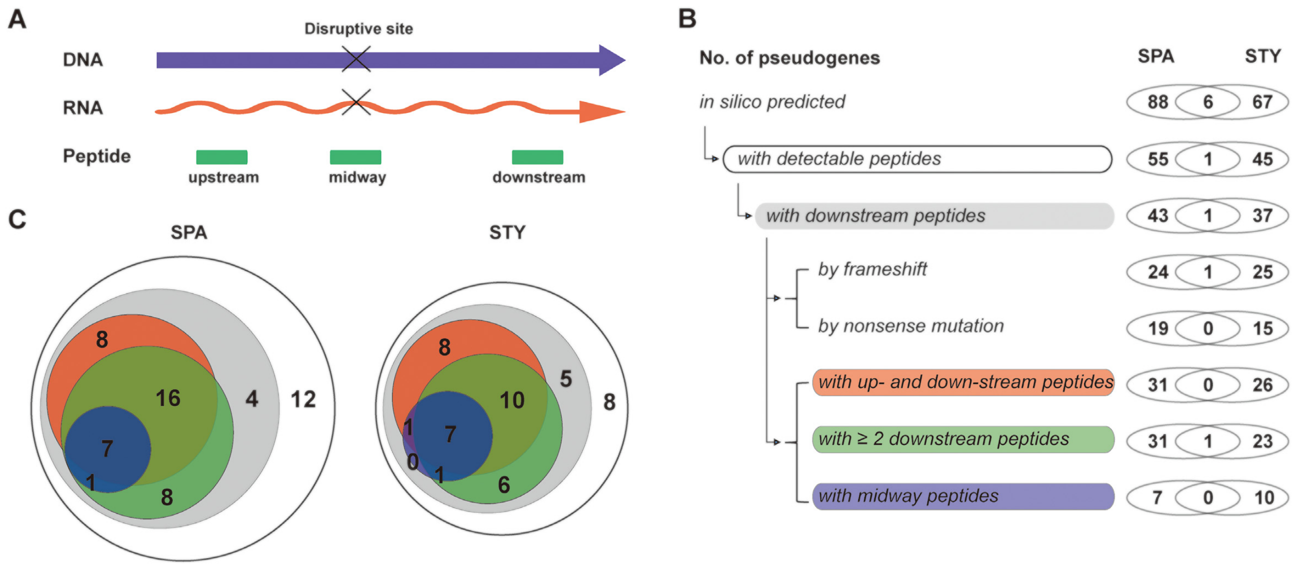

Figure 1. Overview of the expressed pseudogenes in *Salmonella*. **(A)** According to the mapped location, the peptides detected by LC-MS/MS can be divided upstream from the disruptive site, midway (directly upon the disruptive site), and downstream from the disruptive site. **(B)** The pseudogenes are divided into subgroups with different disruptive causes and different levels of supporting evidence by proteomic data. **(C)** Venn diagrams of pseudogenes of different subgroups. The three numbers in the oval shapes (from left to right) represent pseudogenes disrupted in *S.* Paratyphi A only, disrupted in both *S.* Paratyphi A and *S.* Typhi, and disrupted in *S.* Typhi only. The colour assigned is the same as that of the panel B. Abbreviation: SPA, *S.* Paratyphi A; STY, *S.* Typhi.

searches guaranteed that these peptides were gene-specific; i.e. they did not exist in any other genes.

As the sequence upstream of the start codon in *S.* Typhimurium is nearly identical to that in *S.* Paratyphi A or *S.* Typhi, the transcriptional and translational regulatory elements in *S.* Paratyphi A or *S.* Typhi remain intact in theory. Accordingly, the pseudogenes can still act as a template to initialize transcription and translation and express the peptides upstream of the disruptive site, but fail to express the peptides downstream. For the 81 pseudogenes in *S.* Paratyphi A and/or *S.* Typhi, however, the peptides were detected downstream from the disruptive site (Figure 1). We then considered another, albeit less likely, scenario: an alternative start codon downstream from the disruptive site may have initiated the translation process. Nevertheless, 57 pseudogenes had their peptides mapped both upstream and downstream from the disruptive sites (Supplementary Table S4). In addition, there were even 17 pseudogenes with peptides mapped directly onto the disruptive sites (Supplementary Table S5).

To further eliminate the false positive identification caused by similar m/z values when analyzing the mass-spectrometry data, we further screened 55 pseudogenes whose expression was supported by at least two peptides detected downstream or directly upon the disruptive sites (Supplementary Table S6). All the above findings suggested that a considerable portion of pseudogenes in *Salmonella* were indeed able to express full-length proteins.

Pseudogenes are expressed at a low level

Here, the 81 pseudogenes with their peptides mapped downstream or upon the disruptive site were considered truly expressed pseudogenes and were included for the following analyses. To quantify the expression using the abundance of the shared peptides, the pseudogenes were, on average, merely 10% of their intact homologs in *S.* Typhimurium (Figure 2). When we took 2-fold as the threshold of low expression, under both the SPI-1- and SPI-2-mimicking conditions, 30 out of 43 pseudogenes in *S.* Paratyphi A and 30 out of 37 pseudogenes in *S.* Typhi were found to be lowly expressed.

This low expression level can be explained by the following hypotheses: 1) it is difficult for standard transcription/translation machinery to express pseudogenes, and 2) pseudogenes are not required to be expressed. To avoid the latter hypothesis, a further comparison was made between *S.* Paratyphi A and *S.* Typhi. Of the 81 genes, only one was disrupted in both *S.* Paratyphi A and *S.* Typhi; the remaining genes were disrupted in one serovar but remained intact in the other. Still taking two-fold as threshold, for both *S.* Paratyphi A and *S.* Typhi under either SPI-1- or SPI-2-mimicking conditions, the majority of pseudogenes were lowly expressed relative to their intact homologs in the other human-restricted serovar (Figure 2), thereby supporting the former hypothesis.

The expression profile revealed from the transcriptome was inconsistent with that from the proteome (Pearson correlation $R^2 < 0.5$, $P > 0.05$ for both *S.* Paratyphi A and *S.* Typhi). At the RNA level, the expression level of pseudogenes was overall similar to their intact homologs in *S.* Typhi; over half of these pseudogenes under SPI-2-mimicking conditions displayed an even higher expression than their intact homologs (Figure 2D).

We divided the pseudogenes into two categories according to whether they had been disrupted via frameshift or nonsense mutation. Still taking the expression of the intact copy in *S.* Typhimurium as the baseline, the relative expres-

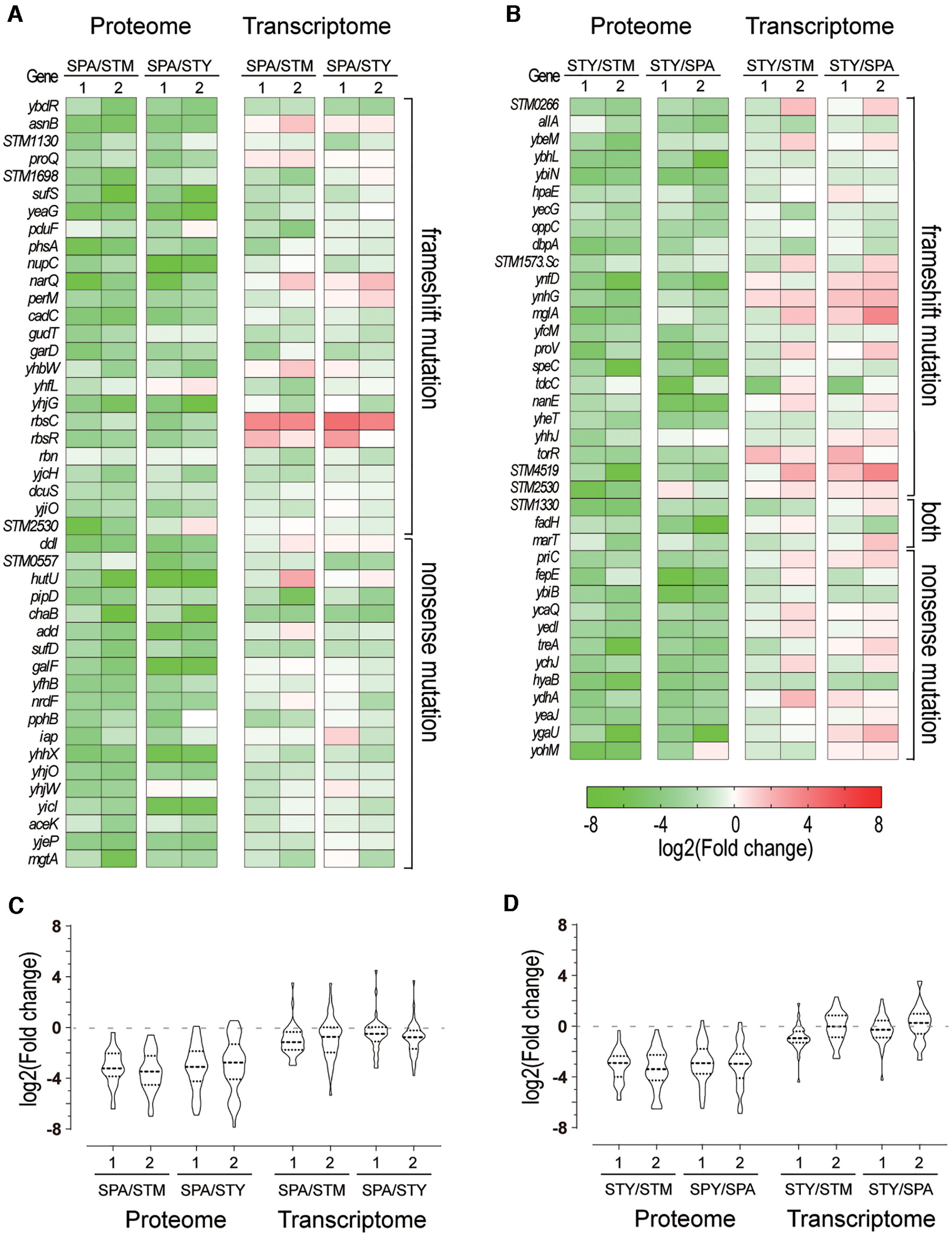

Figure 2. Expression of pseudogenes at the transcriptomic and translational levels for *S.* Paratyphi A and *S.* Typhi. The pseudogenes' relative expression (fold change) is calculated using the expressions of the pseudogenes in *S.* Typhi or *S.* Paratyphi A divided by the expression of their intact homologs in *S.* Typhimurium. The transcriptome and proteome are measured under two conditions: 1, SPI-1 mimicking condition; 2, SPI-2 mimicking condition. The relative expression is shown as heatmap in panel A and B and as violin plot in panel C and D. Panel A and C show the expression of pseudogenes in *S.* Paratyphi A; Panel B and D show the expression of pseudogenes in *S.* Typhi. Abbreviation: SPA, *S.* Paratyphi A; STY, *S.* Typhi; STM, *S.* Typhimurium. Taking 'SPA/STY' for example, it means the genes are disrupted in SPA (the former) but intact in STY (the latter).

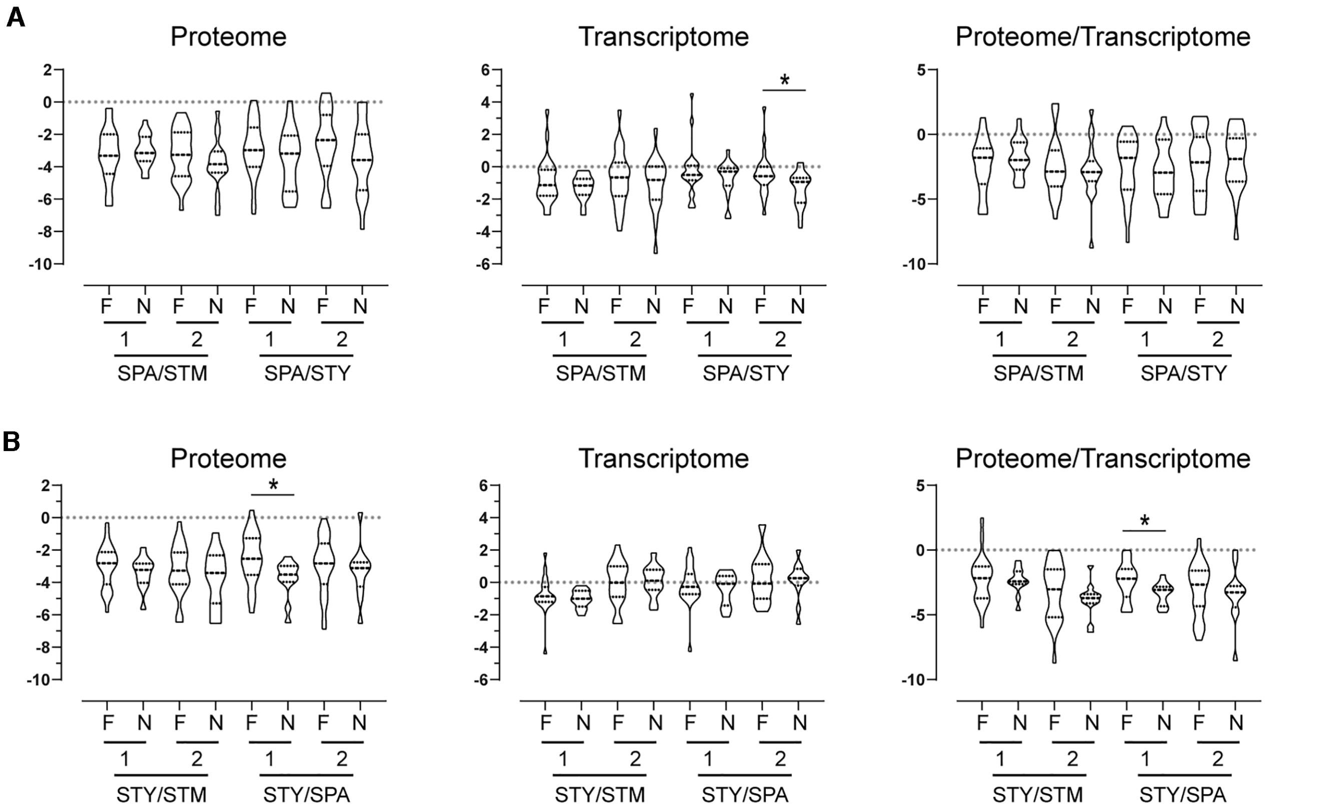

Figure 3. Comparison of expression of pseudogenes disrupted by frameshift and nonsense mutation. **(A)** Expression of pseudogenes from *S.* Paratyphi A. **(B)** Expression of pseudogenes from *S.* Typhi. The fold change of expression shown by y axis is log2 transformed. Abbreviation: F, frameshift; N, nonsense mutation; 1, SPI-1 mimicking condition; 2, SPI-2 mimicking condition; SPA, *S.* Paratyphi A; STY, *S.* Typhi; STM, *S.* Typhimurium. *, $P < 0.05$ (Student's t test).

sions between the two categories did not display constant differences at both the protein and RNA levels (Figure 3). Next, we calculated translational efficiency by dividing protein expression by RNA expression. Still, we did not observe constant differences between the two categories of pseudogenes among the two growth conditions (Figure 3).

Sequence features of recoding

Of the 55 genes whose decoding involved frameshifting, single nucleotide deletion was the most frequent pseudogenization cause (63.6%), followed by single nucleotide insertion (20.0%) (Supplementary Table S7). In sixteen (29.1%) pseudogenes, the disruptive sites were located within mononucleotide repeats, fourteen of which were adenine or thymine (Figure 4). For ten of the 55 genes, we identified a peptide encoded by the frameshift junction flanking sequence, indicating the compensatory frameshifting occurred nearby or even exactly at the mutation site (Supplementary Table S5). Three of them specified glutamine at the frameshift site, and the remaining genes specified other amino acids. Two conserved motifs were found upstream and downstream of the frameshift site of the ten genes. However, since they were not located at the same position or the same frame (Figure 4A), whether or not the two motifs stimulated frameshifting was doubtful.

Of the 35 genes with readthrough of in-frame stop codon, an exceptionally high proportion (68.6%) involved UAG as the stop codon (Figure 4B). In contrast, for the normal genes in the *Salmonella* genome, the percentage of UAG stop codons (9.3%) was significantly lower than that of UAA (61.4%) and UGA (29.3%) (Chi square test, $P < 0.001$). These nonsense mutations were mostly caused by C→T and G→A transitions. Seven genes had a peptide mapped upon the in-frame stop codon. The six UAG codons were translated into four glutamine, one cysteine and one phenylalanine (Supplementary Table S5); the remaining UGA codon, which was in the gene SPA_RS06975, was translated into arginine. Notably, SPA_RS06975 was the only pseudogene in *S.* Typhi and *S.* Paratyphi A which had its in-frame stop codon with a potential selenocysteine insertion sequence (SECIS, see Supplementary Figure S1) (48).

To search for sequence features that stimulate the readthrough of in-frame stop codon, we investigated the surrounding nucleotide composition. At the -1 and + 4 site, which were immediately adjacent to the in-frame stop codon, cytosine accounted for 58.3% and 37.5%, whereas uracil accounted for 0.0% and 12.5% (Figure 4C). Half the bases at the + 5 site were adenine. Compositions at these sites were significantly different between the translated pseudogenes and the normal genes (Chi square test,

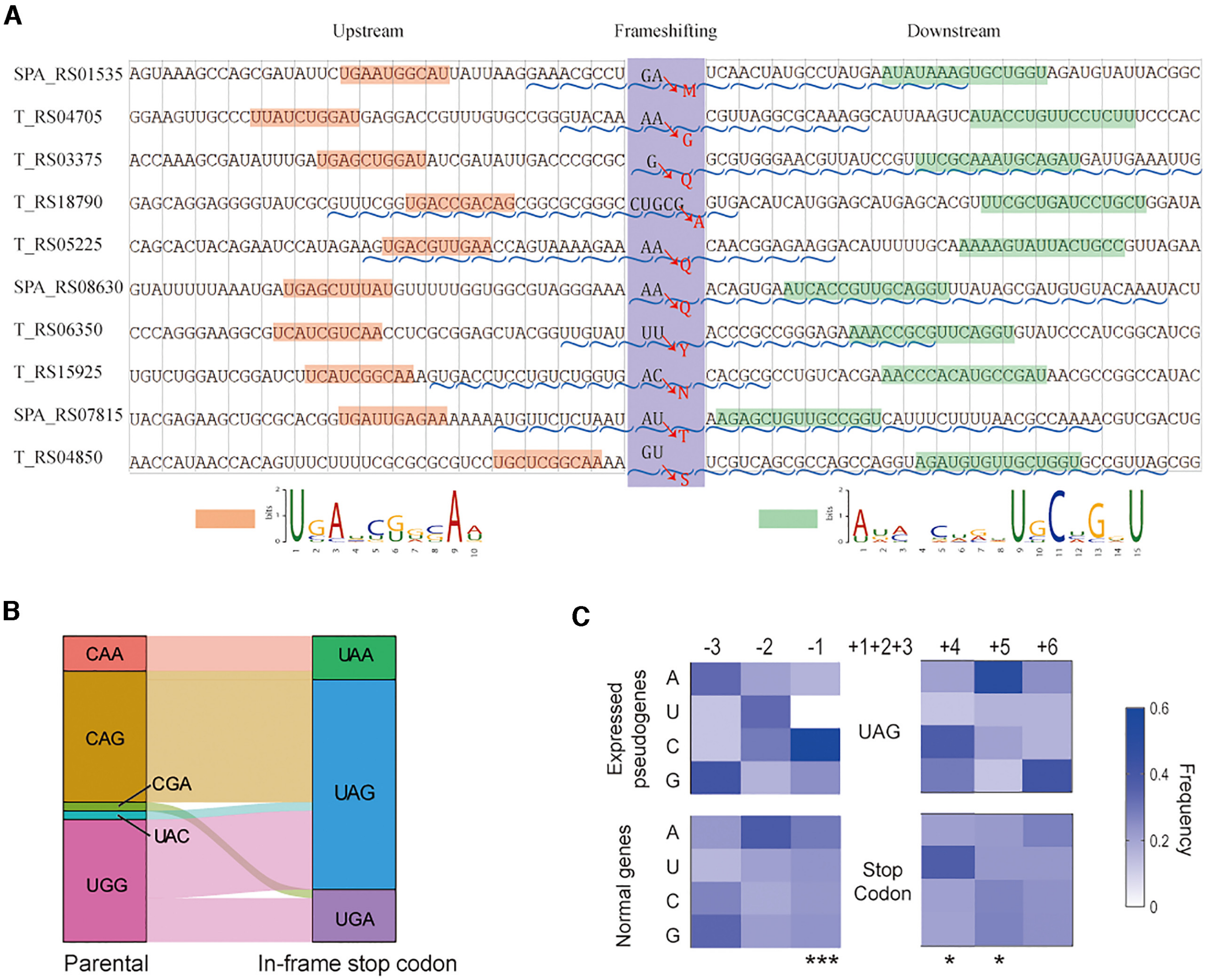

Figure 4. Sequence features surrounding the recoding site. **(A)** Ten genes which have peptides detected upon the frameshift mutation site. The locations of the detected peptides are represented as the wavy lines. The amino acids specified at the frameshift sites, which is detected by LC-MS/MS, are marked as the red letters in the middle. The upstream and downstream conserved motifs, which are searched by MEME software, are represented by the red boxes and the green boxes. The sequence logos of the two motifs are listed at the bottom. **(B)** Nucleotide composition of the in-frame stop codon (after mutation) and their parental codon (before mutation). **(C)** Heatmap showing the nucleotide composition upstream and downstream of the in-frame stop codon. Normal genes are compared as the control. Positions with significant difference between the translated pseudogenes and the normal genes are marked with asterisks. *, $P < 0.05$; ***, $P < 0.001$ (Chi square test).

$P < 0.01$). Nevertheless, we did not observe any significant enrichment at the unit of codon or encoded amino acid.

Pseudogenes in *Salmonella* can express in *E. coli*

To further validate the expressions of pseudogenes and to test if recoding is common among *Enterobacteriaceae*, we individually cloned fifteen pseudogenes of *S.* Typhi and expressed them in *E. coli* (Figure 5, Supplementary Table S8). These genes were chosen because they had at least two identified peptides encoded downstream of the disruptive site and therefore their protein expression was less likely to result from false positive identification of mass spectrometry. Since the fluorescent protein SYFP2 was fused downstream of the pseudogenes, the fluorescence signals could be detected only if the disruptive sites were successfully translated. In contrast to mass spectrometry, which regards the entire bacterial culture as a whole, flow cytometry can measure gene expression at a single-cell level and reveal heterogeneity in a bacterial population. The medium percentage of bacterial cells carrying pseudogenes that released fluorescence signals was 2.6%, as compared to 68.0% for the cells carrying the intact homolog and 0.0% for the negative control (Figure 5, Supplementary Figure S2). Accordingly, the ratio of the cells expressing full-length pseudogenes relative to those expressing their intact homologs ranged between 0.02–0.17. The ratios observed from the flow cytometry were not exactly correlated with that from the LC-MS/MS results under the SPI-1-mimicking conditions (Pearson correlation r = 0.27; $P > 0.05$); overall, the former was lower than the latter (Student's t-test, $P < 0.05$). This discrepancy could be due to certain false-positive LC-MS/MS results, or due to the different bacterial hosts expressing the pseudogenes.

The mRNA-derived cDNA of these pseudogenes was sequenced and found to be identical to the DNA tem-

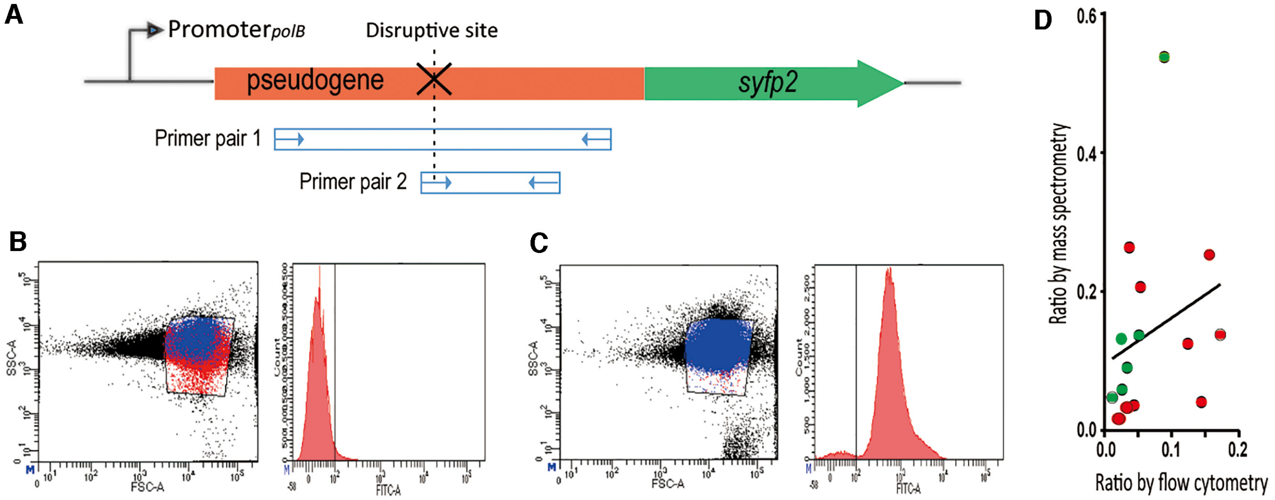

Figure 5. Expression of *Salmonella* pseudogenes in *E. coli*. **(A)** Design of the expression validation for the pseudogenes. The constitutive-expression vector involved *syfp2*, which encodes a fluorescent protein at the 3′end of the query gene and is used for expression quantification by flow cytometry. Two primer pairs are designed for each gene. Primer pair 1 is used to amplify the mRNA-derived cDNA, with the PCR product being sequenced to confirm the presence of mutation at the DNA and RNA level. Primer pair 2 incorporates the nucleotides of the intact homolog at the disruptive sites, which was used to amplify the possible gene products carrying reverse mutations that restore the function of pseudogenes. **(B)** and **(C)** Flow cytometry results of the expressed pseudogenes and the corresponding intact homologs (take the gene *mglA* for example). For the dot plots, the red dots and the blue dots represent bacterial cells with fluorescence intensity lower and higher than 100 (a threshold set based on negative control), respectively. The histograms show the distribution of fluorescence intensity in the bacterial population. **(D)** Ratio of the expressed pseudogenes relative to their intact homologs, which is detected by flow cytometry and mass spectrometry (LC-MS/MS), respectively. Red dots represent pseudogenes disrupted by frameshifts; green dots represent pseudogenes disrupted by nonsense mutation.

plate. We also designed primers that incorporated the nucleotides of the intact homolog at the disruptive sites (Figure 5A). No PCR product was detected, indicating that the expression could hardly have resulted from a tiny frequency of reverse mutations at either the DNA or RNA level. The above findings demonstrated again that the low expression of pseudogenes was achieved by translational recoding.

Out of the fifteen pseudogenes, we further selected four for validation by a second run of mass-spectrometry. The putative product of the pseudogenes were pulled with the fused His-tag out of the total protein lysate. Then the expected peptides were identified by the MALDI-TOF MS (Supplementary Table S9 & Supplementary Figure S3), which validated the protein expression of pseudogenes in *E. coli*.

Virulence is affected by restoring pseudogenes' functions

We have found many pseudogenes can express full-length proteins, yet with a low expression. To investigate whether the virulence can be altered by fully restoring the functions of the pseudogenes, we expressed the intact homologs of pseudogenes in *S.* Typhi with a strong promoter, and infected the Caco-2 and THP-1 cell lines, which mimicked the two hallmarks of *Salmonella* pathogenesis, namely invasion of intestine epithelial cells and replication inside host macrophages (53). When comparing the mutant with the wild strain and taking two-fold difference with statistical significance as the criterion, as many as seven mutants reduced their capability to infect Caco-2, whereas only one mutant enhanced the intracellular survival in THP-1 (Figure 6A,B).

A unique feature of typhoid fever is asymptomatic carriage within the gallbladder, which is linked to shedding of *S.* Typhi from the gastrointestinal tract (54). Thus, we investigated whether pseudogenization could enhance or attenuate the bile tolerance. Only two mutants showed significantly different growth rate in bile-containing medium, yet the differences were both smaller than 5% (Figure 6C).

DISCUSSION

Despite decades of research, pseudogenes are still as obscure and misunderstood as they were when they were first discovered. Previous proteomics studies have discovered a dozen 'pseudo-pseudogenes' in *M. tuberculosis* and *S. glossinidius*, which express full-length proteins through recoding, but many of them are mobile and/or phage-derived (37,39). It is known that recoding is common in virus (24,25), but whether recoding is common for eubacterial genes remains unclear. The current study's *in silico* analysis was restricted to genes conserved throughout *Salmonella* serogroups, rather than mobile genes. For the first time, we discovered that half of pseudogenes in *Salmonella* could express full-length proteins using a combination of shotgun proteomics and a fluorescent reporter system. Due to technical difficulty (e.g. extraction of membrane bound proteins) as well as the limited culturing conditions conducted, the bacterial proteome has not been 100% covered; thus the expressed pseudogenes are still under-represented. The large number of expressed pseudogenes detected in *Salmonella* than that in *M. tuberculosis* and *S. glossinidius* may result from the different age of pseudogenes. *S.* Typhi is a young serovar that has not diverged from its host-generalist ancestor until 50,000 years ago (55). Con-

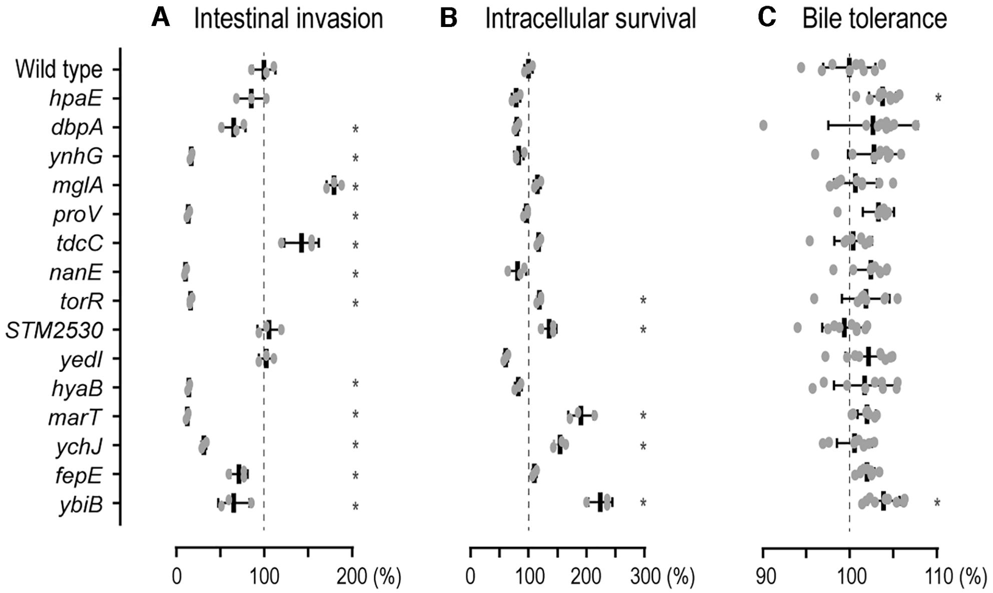

Figure 6. Role of pseudogenization in bacterial fitness and virulence. **(A)** Invasion into Caco-2 epithelial cells. **(B)** Intracellular survival in THP-1 macrophage cells. **(C)** Bacterial growth rate in bile-containing medium. The compared bacteria are *S.* Typhi strain CMCC50079 (Wild strain) and its isogenic mutants that express the intact copy of the fifteen pseudogenes. Data from the mutant strains are presented as the percentages relative to the wild strain and are visualized as scatter plot, with the lines representing mean and SD. *, $P < 0.05$ (One-way ANOVA corrected by Dunnett's multiple comparisons test).

sequently, its pseudogenes are new-born and mostly carry only one disruptive mutation. In contrast, most of pseudogenes in *M. tuberculosis* and *S. glossinidius* have accumulated multiple disruptive mutations, for which complete recoding is much more difficult.

The constitutionally low expression of the pseudogenes seems independent of the growth conditions and bacterial genetic backgrounds. At the RNA level, the expression of pseudogenes can be comparable to their intact homologs. This is not a surprise since, in the genome of *S.* Typhi and *S.* Paratyphi A, the promoter regions and ribosomal binding sites have not been altered, which still initiate regular transcription and translation. Furthermore, the mRNA sequences of the expressed pseudogenes made little difference to their DNA templates. The disrupted state at the RNA level demonstrated that transcriptional correction was not the mainstream approach for eubacterial recoding.

Recoding through ribosomal frameshifting and codon redefinition were both detected in the present study. Majority of the frameshifting occurred on -1-frameshift mutations, but the detailed mechanism of frameshifting remains unclear. For codon redefinition, the UGA codon has once been described as particularly leaky and can be read as tryptophan at an appreciable frequency (24,56). This recoding results from occasional recognition of the stop codon by the EF-Tu:Trp-tRNA$^{\mathrm{Trp}}$:GTP ternary complex, which competes with release factor 2 that releases the elongating peptide (57). But in the present study, UAG is the most abundant translated stop codon, whereas in the normal genes it accounts for < 10%. The non-essentiality of UAG in *Salmonella* makes it more likely that inefficient recognition of UAG by release factor 1 facilitates decoding by near-cognate aminoacyl tRNA and so of readthrough. Glutamine instead of tryptophan seems to be the most likely amino acid recruited by UAG; this is probably because tRNA$^{\mathrm{Gln}}$ that recognizes CAG mis-pairs with UAG. The exact biochemical basis may be different between the recoding of UAG and UGA because of their different codon context. Adenine is found to be immediately adjacent to the 3′ side of UGA (58); but UAG is surrounded by cytosine.

Through flow cytometry we found only a subset of bacterial population can express pseudogenes. This heterogeneity has also been reported in viral decoding. With live-cell single RNA imaging technology, Lyon *et al.* reported 8% of HIV-1 RNA molecules frameshifted for translation *in vivo*; the frameshifting was not continuous but occurred in bursts on single RNA that can last for several rounds of translation (59). They also found that, regardless of external conditions, the fraction of frameshifting RNA remained constant. This is consistent with our finding that the expression level of pseudogenes in SPI-1- and SPI-2-mimicking media is similar to each other. It is therefore likely that, recoding efficiency depends only on pseudogene sequence and translational apparatus.

Pseudogenization is thought to result from gene redundancy during switch of niche or from the adaptive advantage brought by gene loss. We found that pseudogenization of *ynhG*, *proV*, *nanE*, *torR*, *hyaB*, *marT* and *ychJ* can favor invading epithelium cell and entering the reticuloendothelial system, which was the same as the previously reported *sopE2* and *sseJ* (19,20). We also found that pseudogenization of *ybiB* can favor survival in human macrophages by decreasing intracellular bacterial replication, which was the same as *sopA* (60). If pseudogenization of certain genes contributes to *S.* Typhi's adaptability in human systemic infection, why do these pseudogenes nonetheless yield low protein expression through recoding? Is the recoding merely a translational error? Another paradox is that, since pseudogenes are non-functional and hence costly to the organism, why are they persistent in bacterial genomes? The mean half-life of pseudogenes in *Buchnera*, an endocellular bacterial symbiont found in insects, is approximately 24 million years (61). Moreover, the dN/dS values (ratios of substitution rates at non-synonymous and synonymous sites) of pseudogenes are significantly lower than 1 (62), indicative of a strong purifying selection on pseudogenes, namely pseudogenes are still functional to some extent.

We hereby propose the 'Compensation theory' as an explanation. The process of host adaptation has not fully finished for the host-adapted *Salmonella* serovars; the relevant genes and pathways may have evolved at an intermediate stage. Once the bacteria revert to a free-living state or occasionally infect non-specific hosts, the functions that are dispensable or adaptively obligatory in the specific host would be required again. To this end bacteria need a switch to control gene expression, which can be accomplished through pseudogene recoding (at the protein level), replication slippage (at the DNA level) or transcription slippage (at the RNA level). These mechanisms are often mediated by simple sequence repeats (SSR). SSR is therefore termed 'contingency loci' because they allow bacteria to undergo a reversible and inheritable phenotypic shift that helps bacteria adapt to environment, such as facilitating evasion of a host's immune defense (63). The 'Compensation theory' may explain why *S.* Paratyphi A and *S.* Typhi have only a few overlapping pseudogenes: the genes found to be disrupted in one serovar but intact in the other are probably still viable. Their pseudogenes would not fully overlap until the two young human-specialist serovars become completely adapted to their host.

Notably, the functional compensation may require only a minimum protein expression that recoding of pseudogene is already affordable. For example, the Tk gene in Herpes simplex virus (HSV) encodes thymidine kinase that is necessary for viral replication. Meanwhile, Tk is the target of the drug acyclovir so that HSV often inactivates its Tk gene through frameshift to derive drug resistance. Through recoding the Tk mutant can still produce as much as 3% of full-length thymidine kinase in the wild strain (64). Such a small amount is already biologically relevant because it suffices to permit viral replication (65).

CONCLUSIONS

The most important finding of the present study is that majority of the eubacterial pseudogenes in *Salmonella* may still possess a low protein-coding potential. By cellular infection experiments, we demonstrated that whether or not these pseudogenes regain their functionality affects virulence, which further highlights the plasticity of translational recoding acting as a phenotypic buffer. As a crucial complement to the standard central dogma, understanding the recoding mechanism can shed light on yet unknown details of translation machinery. To uncover DNA motifs that promote recoding, identification of complete sets of pseudogenes as training data is critically needed. In the future, an 'omic'-level reporter system, in conjunction with genomic, transcriptomic and proteomic techniques, will be able to provide a more accurate annotation and capture the real complexity of the translation landscape.

DATA AVAILABILITY

The transcriptome data have been deposited to NCBI SRA database with the BioProject PRJNA638315. The LC-MS/MS data have been deposited to iProX database (https://www.iprox.org/) with the ProteomeXchange dataset identifier PXD022254. The flow cytometry data have been deposited to FlowRepository database under Experiment_ID FR-FCM-Z4PH.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

The study was supported by the National Natural Science Foundation, China (grant number 82072241 and 31670132 to Y. Feng) and Chang Gung Memorial Hospital, Taiwan (CMRPG3G1453, CMRPG3L0541, and CMRPD1L0031 to C.-H. Chiu).
Conflict of interest statement. None declared.

REFERENCES

1. Lerat,E. and Ochman,H. (2005) Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res.*, **33**, 3125–3132.
2. Goodhead,I. and Darby,A.C. (2015) Taking the pseudo out of pseudogenes. *Curr. Opin. Microbiol.*, **23**, 102–109.
3. Bobay,L.M. and Ochman,H. (2017) The evolution of bacterial genome architecture. *Front. Genet.*, **8**, 72.
4. Zhang,Z.D., Frankish,A., Hunt,T., Harrow,J. and Gerstein,M. (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol.*, **11**, R26.
5. Frankish,A., Diekhans,M., Jungreis,I., Lagarde,J., Loveland,J.E., Mudge,J.M., Sisu,C., Wright,J.C., Armstrong,J., Barnes,I. *et al.* (2021) Gencode 2021. *Nucleic Acids Res.*, **49**, D916–D923.
6. Lo,W.S., Huang,Y.Y. and Kuo,C.H. (2016) Winding paths to simplicity: genome evolution in facultative insect symbionts. *FEMS Microbiol. Rev.*, **40**, 855–874.
7. Toh,H., Weiss,B.L., Perkin,S.A., Yamashita,A., Oshima,K., Hattori,M. and Aksoy,S. (2006) Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of sodalis glossinidius in the tsetse host. *Genome Res.*, **16**, 149–156.
8. Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Garnier,T., Churcher,C., Harris,D. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.
9. Kuo,C.H. and Ochman,H. (2010) The extinction dynamics of bacterial pseudogenes. *PLos Genet.*, **6**, e1001050.
10. Alikhan,N.F., Zhou,Z., Sergeant,M.J. and Achtman,M. (2018) A genomic overview of the population structure of salmonella. *PLos Genet.*, **14**, e1007261.
11. Kurtz,J.R., Goggins,J.A. and McLachlan,J.B. (2017) Salmonella infection: interplay between the bacteria and host immune system. *Immunol. Lett.*, **190**, 42–50.
12. Tanner,J.R. and Kingsley,R.A. (2018) Evolution of salmonella within hosts. *Trends Microbiol.*, **26**, 986–998.
13. Chiu,C.H., Tang,P., Chu,C., Hu,S., Bao,Q., Yu,J., Chou,Y.Y., Wang,H.S. and Lee,Y.S. (2005) The genome sequence of salmonella enterica serovar choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res.*, **33**, 1690–1698.

14. Holt,K.E., Thomson,N.R., Wain,J., Langridge,G.C., Hasan,R., Bhutta,Z.A., Quail,M.A., Norbertczak,H., Walker,D., Simmonds,M. *et al.* (2009) Pseudogene accumulation in the evolutionary histories of salmonella enterica serovars paratyphi a and typhi. *BMC Genomics*, **10**, 36.
15. Liu,W.Q., Feng,Y., Wang,Y., Zou,Q.H., Chen,F., Guo,J.T., Peng,Y.H., Jin,Y., Li,Y.G., Hu,S.N. *et al.* (2009) Salmonella paratyphi C: genetic divergence from salmonella choleraesuis and pathogenic convergence with salmonella typhi. *PLoS One*, **4**, e4510.
16. Spano,S. and Galan,J.E. (2012) A Rab32-dependent pathway contributes to salmonella typhi host restriction. *Science*, **338**, 960–963.
17. Thomson,N.R., Clayton,D.J., Windhorst,D., Vernikos,G., Davidson,S., Churcher,C., Quail,M.A., Stevens,M., Jones,M.A., Watson,M. *et al.* (2008) Comparative genome analysis of salmonella enteritidis PT4 and salmonella gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res.*, **18**, 1624–1637.
18. Matthews,T.D., Schmieder,R., Silva,G.G., Busch,J., Cassman,N., Dutilh,B.E., Green,D., Matlock,B., Heffernan,B., Olsen,G.J. *et al.* (2015) Genomic comparison of the closely-related salmonella enterica serovars enteritidis, dublin and gallinarum. *PLoS One*, **10**, e0126883.
19. Valenzuela,L.M., Hidalgo,A.A., Rodriguez,L., Urrutia,I.M., Ortega,A.P., Villagra,N.A., Paredes-Sabja,D., Calderon,I.L., Gil,F., Saavedra,C.P. *et al.* (2015) Pseudogenization of sopA and sopE2 is functionally linked and contributes to virulence of salmonella enterica serovar typhi. *Infect. Genet. Evol.*, **33**, 131–142.
20. Trombert,A.N., Berrocal,L., Fuentes,J.A. and Mora,G.C. (2010) S. Typhimurium sseJ gene decreases the s. Typhi cytotoxicity toward cultured epithelial cells. *BMC Microbiol.*, **10**, 312.
21. Baranov,P.V., Gurvich,O.L., Fayet,O., Prere,M.F., Miller,W.A., Gesteland,R.F., Atkins,J.F. and Giddings,M.C. (2001) RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. *Nucleic Acids Res.*, **29**, 264–267.
22. Gesteland,R.F. and Atkins,J.F. (1996) Recoding: dynamic reprogramming of translation. *Annu. Rev. Biochem.*, **65**, 741–768.
23. Prieto-Godino,L.L., Rytz,R., Bargeton,B., Abuin,L., Arguello,J.R., Peraro,M.D. and Benton,R. (2016) Olfactory receptor pseudo-pseudogenes. *Nature*, **539**, 93–97.
24. Parker,J. (1989) Errors and alternatives in reading the universal genetic code. *Microbiol. Rev.*, **53**, 273–298.
25. Atkins,J.F., Loughran,G., Bhatt,P.R., Firth,A.E. and Baranov,P.V. (2016) Ribosomal frameshifting and transcriptional slippage: from genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.*, **44**, 7007–7078.
26. Huseby,D.L., Brandis,G., Alzrigat,Praski and Hughes,D. (2020) Antibiotic resistance by high-level intrinsic suppression of a frameshift mutation in an essential gene. *Proc. Nat. Acad. Sci. USA*, **117**, 3185–3191.
27. Javid,B., Sorrentino,F., Toosky,M., Zheng,W., Pinkham,J.T., Jain,N., Pan,M., Deighan,P. and Rubin,E.J. (2014) Mycobacterial mistranslation is necessary and sufficient for rifampicin phenotypic resistance. *Proc. Nat. Acad. Sci. USA*, **111**, 1132–1137.
28. Dohrmann,P.R., Correa,R., Frisch,R.L., Rosenberg,S.M. and McHenry,C.S. (2016) The DNA polymerase III holoenzyme contains gamma and is not a trimeric polymerase. *Nucleic Acids Res.*, **44**, 1285–1297.
29. Urrutia,I.M., Fuentes,J.A., Valenzuela,L.M., Ortega,A.P., Hidalgo,A.A. and Mora,G.C. (2014) Salmonella typhi shdA: pseudogene or allelic variant?*Infect. Genet. Evol.*, **26**, 146–152.
30. Xu,J., Hendrix,R.W. and Duda,R.L. (2004) Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol. Cell*, **16**, 11–21.
31. Antonov,I., Coakley,A., Atkins,J.F., Baranov,P.V. and Borodovsky,M. (2013) Identification of the nature of reading frame transitions observed in prokaryotic genomes. *Nucleic Acids Res.*, **41**, 6514–6530.
32. Larsen,B., Wills,N.M., Nelson,C., Atkins,J.F. and Gesteland,R.F. (2000) Nonlinearity in genetic decoding: homologous DNA replicase genes use alternatives of transcriptional slippage or translational frameshifting. *Proc. Nat. Acad. Sci. USA*, **97**, 1683–1688.
33. Baranov,P.V., Fayet,O., Hendrix,R.W. and Atkins,J.F. (2006) Recoding in bacteriophages and bacterial IS elements. *Trends Genet.*, **22**, 174–181.
34. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2002) Release factor 2 frameshifting sites in different bacteria. *EMBO Rep.*, **3**, 373–377.
35. Sharma,V., Firth,A.E., Antonov,I., Fayet,O., Atkins,J.F., Borodovsky,M. and Baranov,P.V. (2011) A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. *Mol. Biol. Evol.*, **28**, 3195–3211.
36. Schrimpe-Rutledge,A.C., Jones,M.B., Chauhan,S., Purvine,S.O., Sanford,J.A., Monroe,M.E., Brewer,H.M., Payne,S.H., Ansong,C., Frank,B.C. *et al.* (2012) Comparative omics-driven genome annotation refinement: application across yersiniae. *PLoS One*, **7**, e33903.
37. Bespyatykh,J., Smolyakov,A., Guliaev,A., Shitikov,E., Arapidi,G., Butenko,I., Dogonadze,M., Manicheva,O., Ilina,E., Zgoda,V. *et al.* (2019) Proteogenomic analysis of mycobacterium tuberculosis beijing B0/W148 cluster strains. *J. Proteomics*, **192**, 18–26.
38. Gupta,N., Tanner,S., Jaitly,N., Adkins,J.N., Lipton,M., Edwards,R., Romine,M., Osterman,A., Bafna,V., Smith,R.D. *et al.* (2007) Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res.*, **17**, 1362–1377.
39. Goodhead,I., Blow,F., Brownridge,P., Hughes,M., Kenny,J., Krishna,R., McLean,L., Pongchaikul,P., Beynon,R. and Darby,A.C. (2020) Large-scale and significant expression from pseudogenes in sodalis glossinidius - a facultative bacterial endosymbiont. *Microb. Genom.*, **6**, e000285.
40. Page,A.J., Cummins,C.A., Hunt,M., Wong,V.K., Reuter,S., Holden,M.T., Fookes,M., Falush,D., Keane,J.A. and Parkhill,J. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691–3693.
41. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and genomewise. *Genome Res.*, **14**, 988–995.
42. Manes,N.P., Gustin,J.K., Rue,J., Mottaz,H.M., Purvine,S.O., Norbeck,A.D., Monroe,M.E., Zimmer,J.S., Metz,T.O., Adkins,J.N. *et al.* (2007) Targeted protein degradation by salmonella under phagosome-mimicking culture conditions investigated using comparative peptidomics. *Mol. Cell. Proteomics*, **6**, 717–727.
43. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**, 357–359.
44. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.*, **12**, 323.
45. Thorvaldsdottir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinf.*, **14**, 178–192.
46. Tjaden,B. (2020) A computational system for identifying operons based on RNA-seq data. *Methods*, **176**, 62–70.
47. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
48. Zhang,Y. and Gladyshev,V.N. (2005) An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes. *Bioinformatics*, **21**, 2580–2589.
49. Temnykh,S., DeClerck,G., Lukashova,A., Lipovich,L., Cartinhour,S. and McCouch,S. (2001) Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.*, **11**, 1441–1452.
50. Hua,X., Zhang,L., Moran,R.A., Xu,Q., Sun,L., van Schaik,W. and Yu,Y. (2020) Cointegration as a mechanism for the evolution of a KPC-producing multidrug resistance plasmid in proteus mirabilis. *Emerg. microbes & infect.*, **9**, 1206–1218.
51. He,J., Sun,L., Zhang,L., Leptihn,S., Yu,Y. and Hua,X. (2021) A novel SXT/R391 integrative and conjugative element carries two copies of the blaNDM-1 gene in proteus mirabilis. *Msphere*, **6**, e0058821.
52. Shi,L., Ansong,C., Smallwood,H., Rommereim,L., McDermott,J.E., Brewer,H.M., Norbeck,A.D., Taylor,R.C., Gustin,J.K., Heffron,F. *et al.* (2009) Proteome of salmonella enterica serotype typhimurium grown in a low Mg/pH medium. *J. Proteomics Bioinform.*, **2**, 388–397.
53. Johnson,R., Mylona,E. and Frankel,G. (2018) Typhoidal salmonella: distinctive virulence factors and pathogenesis. *Cell. Microbiol.*, **20**, e12939.

54. Crawford,R.W., Rosales-Reyes,R., Ramirez-Aguilar Mde,L., Chapa-Azuela,O., Alpuche-Aranda,C. and Gunn,J.S. (2010) Gallstones play a significant role in salmonella spp. gallbladder colonization and carriage. *Proc. Nat. Acad. Sci. USA*, **107**, 4353–4358.
55. Baker,S. and Dougan,G. (2007) The genome of salmonella enterica serovar typhi. *Clin. Infect. Dis.*, **45**, S29–S33.
56. Roth,J.R. (1970) UGA nonsense mutations in salmonella typhimurium. *J. Bacteriol.*, **102**, 467–475.
57. Fan,Y., Evans,C.R., Barber,K.W., Banerjee,K., Weiss,K.J., Margolin,W., Igoshin,O.A., Rinehart,J. and Ling,J. (2017) Heterogeneity of stop codon readthrough in single bacterial cells and implications for population fitness. *Mol. Cell*, **67**, 826–836.
58. Engelberg-Kulka,H. (1981) UGA suppression by normal tRNA trp in escherichia coli: codon context effects. *Nucleic Acids Res.*, **9**, 983–991.
59. Lyon,K., Aguilera,L.U., Morisaki,T., Munsky,B. and Stasevich,T.J. (2019) Live-Cell single RNA imaging reveals bursts of translational frameshifting. *Mol. Cell*, **75**, 172–183.
60. Ma,S., Liu,X., Ma,S. and Jiang,L. (2021) SopA inactivation or reduced expression is selected in intracellular salmonella and contributes to systemic salmonella infection. *Res. Microbiol.*, **172**, 103795.
61. Gomez-Valero,L., Latorre,A. and Silva,F.J. (2004) The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont buchnera aphidicola. *Mol. Biol. Evol.*, **21**, 2172–2181.
62. Belinky,F., Ganguly,I., Poliakov,E., Yurchenko,V. and Rogozin,I.B. (2021) Analysis of stop codons within prokaryotic protein-coding genes suggests frequent readthrough events. *Int. J. Mol. Sci.*, **22**, 1876.
63. Moxon,R., Bayliss,C. and Hood,D. (2006) Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.*, **40**, 307–333.
64. Pan,D. and Coen,D.M. (2012) Net -1 frameshifting on a noncanonical sequence in a herpes simplex virus drug-resistant mutant is stimulated by nonstop mRNA. *Proc. Nat. Acad. Sci. USA*, **109**, 14852–14857.
65. Besecker,M.I., Furness,C.L., Coen,D.M. and Griffiths,A. (2007) Expression of extremely low levels of thymidine kinase from an acyclovir-resistant herpes simplex virus mutant supports reactivation from latently infected mouse trigeminal ganglia. *J. Virol.*, **81**, 8356–8360.