BRIEF COMMUNICATION
https://doi.org/10.1038/s41564-018-0340-2
nature microbiology

Pseudogene repair driven by selection pressure applied in experimental evolution

Amitesh Anand[1], Connor A. Olson[1], Laurence Yang[1], Anand V. Sastry[1], Edward Catoiu[1], Kumari Sonal Choudhary[1], Patrick V. Phaneuf[1], Troy E. Sandberg[1], Sibei Xu[1], Ying Hefner[1], Richard Szubin[1], Adam M. Feist[1,2] and Bernhard O. Palsson[1,2*]

Pseudogenes represent open reading frames that have been damaged by mutations, rendering the gene product non-functional. Pseudogenes are found in many genomes and are not always eliminated, even if they are potentially 'wasteful'. This raises a fundamental question about their prevalence. Here we report pseudogene *efeU* repair that restores the iron uptake system of *Escherichia coli* under a designed selection pressure during adaptive laboratory evolution.

Iron metabolism is critical for bacterial physiology and pathogenicity[1]. Paradoxically, the availability of this element is limited due to the poor solubility of ferric ion. Microbial genomes often encode several iron uptake systems to mitigate iron deficiency stress[2]. The responses include uptake of elemental iron, exploitation of xenosiderophores, synthesis of endogenous siderophores and minimization of use of iron-costly pathways. Such elaborate metabolic flexibility supports microbial growth in diverse microhabitats and provides a competitive advantage.

Interestingly, laboratory acclimatization has compromised iron acquisition pathways in bacteria[2]. Several *Escherichia coli* strains have lost functionality of the aerobic elemental iron transporter due to a base deletion in the *efeU* gene that results in two open reading frames (ORFs) transcribing non-functional truncated peptides (Fig. 1a)[3]. We performed a detailed analysis of 12,117 *E. coli* genomes from the Pathosystems Resource Integration Center[4] (PATRIC) to examine the distribution and pattern of fragmentation of this gene. We found 234 unique genomes in which *efeU* was fragmented into two ORFs. More than 68% of the 234 genomes had a shorter fragment of 120 base pairs (bp) and a longer fragment of 720 bp (Fig. 1b). Despite having multiple possible ways of bringing a premature stop codon in frame, we observed a clear bias in the pattern of gene fragmentation, suggestive of a preferred gene inactivation mechanism, and we ruled out the possibility of shared ancestry responsible for this observation (Supplementary Fig. 1). This pseudogene has been associated with the laboratory acclimatization of *E. coli*[3]. To examine this association, we identified 169 genomes with two inactive *efeU* fragments that had an annotated 'isolation source'. We could identify 13 genomes reported as laboratory strains, of which 11 possessed the inactive *efeU*, which was a significant enrichment (hypergeometric test, $P < 0.001$). This substantiated the niche-specific appearance of the *efeU* pseudogene and suggests an altered iron metabolism in controlled environments. This reduction of metabolic flexibility motivated us to design a selection pressure on *efeU* that could be implemented in a laboratory evolution experiment to determine if the pseudogene can be repaired[5].

EfeU along with the products of two downstream operonic genes, EfeO and EfeB, form the elemental iron uptake system. The cryptic EfeUOB system of *E. coli* K-12 leaves siderophore-mediated iron acquisition as the major import system in an aerobic environment. To create iron import stress, we knocked out *entC*, the isochorismate synthase involved in the biosynthesis of endogenous siderophore enterobactin (Supplementary Fig. 2). The *entC* knockout strain showed only a minor growth defect (Supplementary Table 1). The *E. coli* genome contains another isochorismate synthase, *menF*, which is involved in the biosynthesis of naphthoquinones (Supplementary Fig. 2). We generated a double knockout strain, Δ*menF*Δ*entC*, to prevent any possibility of repurposing of these isozymes. We also generated a triple knockout strain, Δ*menF*Δ*entC*Δ*ubiC*, by knocking out chorismate lyase (*ubiC*), which is involved in the biosynthesis of ubiquinone, the major respiratory quinone (Supplementary Fig. 2). The triple knockout strain was designed to resolve the impact of iron limitation from any electron transport chain defect.

Both Δ*menF*Δ*entC* and Δ*menF*Δ*entC*Δ*ubiC* could grow well on complex lysogeny broth medium but failed to grow on minimal salts (M9) medium containing ferric chloride (25 μM $FeCl_3$). The addition of an exogenous siderophore, citrate, to the M9 medium supported the growth of Δ*menF*Δ*entC* (Supplementary Fig. 3a) and, thus, established the deficient siderophore responsible for the growth failure. In an attempt to improve the growth rate under this selective pressure, we performed adaptive laboratory evolution (ALE) of these deletion strains. One of the six independent replicates of the double knockout strain and three of the six replicates of the triple knockout strain started to grow robustly (Fig. 1c, Supplementary Fig. 4a). The double knockout strain improved its growth rate to that of the wild-type strain[6]. A significant difference in the growth rate of the evolved double and triple knockout strains supported our understanding that the electron transport chain defect in aerobic conditions results only when the ubiquinone biosynthetic pathway is perturbed. We evolved two double knockout replicates on an iron-rich medium. As anticipated, exogenous iron supplementation restored the growth (Fig. 1c). These results left us with two distinct sets of evolved strains with potentially variable iron uptake abilities.

Whole genome resequencing was performed to reveal the genetic basis of differential ALE outcomes. We discovered that all the replicates that were able to grow without iron supplementation contained insertions or deletions in the *efeU* gene (Fig. 2a, Supplementary Table 2). The Δ*menF*Δ*entC* replicate ALE_39.91 that evolved to grow without extra iron supplementation had a four-base insertion

[1]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. [2]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark. *e-mail: palsson@ucsd.edu

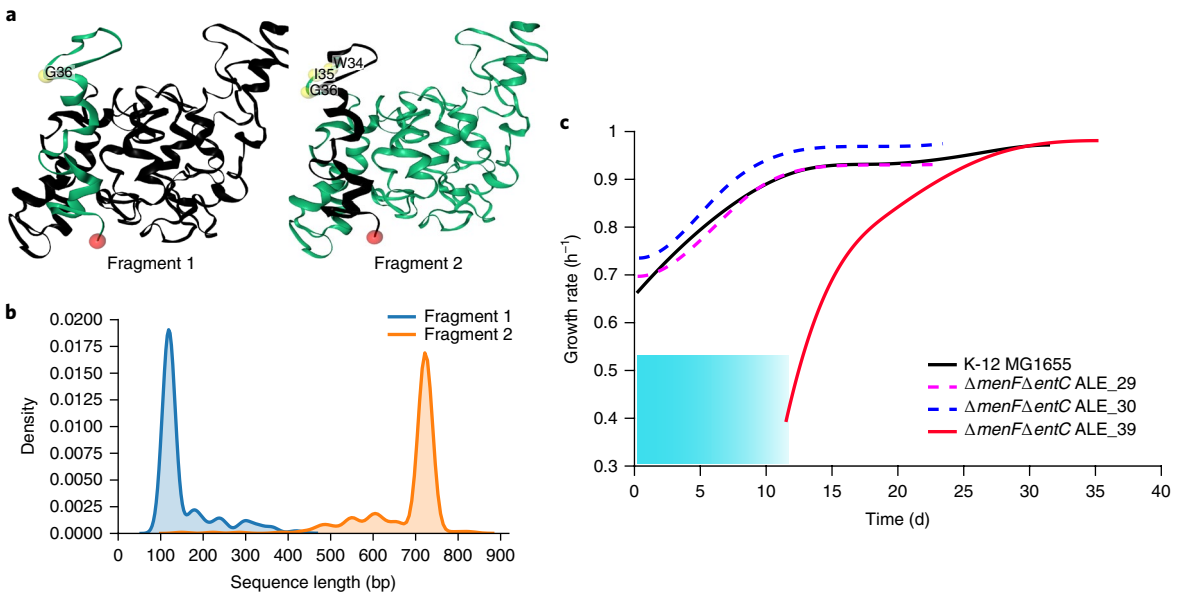

Fig. 1 | The fragmentation and repair of *efeU*. **a**, Overlay of fragmented EfeU peptides on a homology model of functional EfeU. Black indicates deletion; yellow single-nucleotide polymorphism; green identical sequence; and red amino acid residue number 1. **b**, Pattern of *efeU* inactivation across *E. coli* strains. **c**, Evolution trajectory of Δ*menF*Δ*entC*; trajectories with broken and solid lines are for the strains evolved with and without ferrous citrate supplementation, respectively. Evolution of Δ*menF*Δ*entC* without the supplement was started with six independent replicates. The shaded area depicts a zone of no detectable growth rate. The *E. coli* K-12 MG1655 trajectory has been used as a reference.

upstream of the frameshifting deletion that resulted in pseudogenic *efeU*. This insertion restored the complete ORF and the functional form of the protein (Fig. 2b, Supplementary Table 3). The introduction of the same insertion in the parent strain enabled it to grow without siderophores (Supplementary Fig. 3a). Non-functional proteins are a burden to the cellular proteome and an optimization of the cellular proteome for a high growth rate to downregulate the expression of the fragmented *efeU* gene was observed (Fig. 2c). Similar trends were observed in all the Δ*menF*Δ*entC* strains except the strain with the repaired *efeU* gene: ALE_39.91 showed a significant increase in the expression of the *efeUOB* operon (Fig. 2c), indicating the functional significance of this import system. All three independently evolved Δ*menF*Δ*entC*Δ*ubiC* replicates showed similar frame restoration and expression changes in the *efeUOB* operon (Fig. 2a–c, Supplementary Fig. 4b and Supplementary Tables 2 and 3). Interestingly, one replicate showed a same-frame restoring GTAC insertion as observed in the evolved double knockout strain and another showed a similar four-base (CGAG) insertion (Fig. 2a). In all these cases the insertion sequence repeats the upstream bases, which could result from the slippage of the replication complex[7].

The *efeU*-repaired strain could grow robustly with either ferrous or ferric ions (Supplementary Fig. 3a,b), like the *efeU*-based import system reported in *Bacillus subtilis*[8]. The ferric uptake regulator (Fur) is responsible for the precise control of the level of iron in the cytoplasm[9] and therefore we examined the transcriptional status of genes in this regulon. There was a significant decrease in the expression level of genes regulated by Fur compared with the wild-type strain (Supplementary Fig. 5), indicating sufficient intracellular iron levels. This transcriptional rewiring shows the physiological significance of the repaired *efeU*. The alignment of the *efeU* sequences from 11,792 *E. coli* genome IDs with one annotated *efeU* gene revealed genome ID 562.11502, which has an adenine deletion at position 286 and a cytosine insertion at position 328. This indicates a potential natural compensatory frameshift by deletion followed by insertion or vice versa (Supplementary Tables 4 and 5).

Apart from a few cases where pseudogene variants have been implicated in regulatory roles, pseudogenes are classically believed to be insignificant to cellular physiology[10,11]. Our report of the reversion of pseudogenes to their functional states highlights their importance as an 'adaptive repertoire' of selectable traits, even though bacteria display a bias towards elimination of genes that have lost function[12,13]. Furthermore, we demonstrate the use of ALE and designed selection pressures as an experimental approach to ask fundamental genetic questions. The results may alter the view that pseudogenes should be ignored and considered as 'genomic junk.'

Methods

Materials. *E. coli* K-12 MG1655 (American Type Culture Collection (ATCC) 700926) was used as the starting strain. Knockout strains were generated using the P1 phage transduction method[14]. Keio collection strains served as the donor strain for the generation of gene knockout cassettes containing a kanamycin resistance marker[15]. Knockouts were confirmed by PCR and DNA resequencing (PCR confirmation primers are given in Supplementary Table 6). DNA knock-in was performed by pKD46-mediated homologous recombination[16] of DNA amplified using PCR and M9-agar plates were used as selection plates. Recombinants were screened by PCR using positive-recombination-specific primers and Sanger DNA sequencing (PCR confirmation primers are given in Supplementary Table 6). Growth curve analysis was performed using the Bioscreen C Reader system with 200 µl of culture per well. Four biological replicates were used in the assay. Media components were purchased from Sigma. Iron supplementation was performed using one of the following: (1) 20 µM $FeCl_2$; (2) 20 µM $FeCl_3$; and (3) 20 µM $FeCl_3$ and 10 mM sodium citrate.

Analysis of *efeU* across *E. coli* strains. We downloaded genome metadata from PATRIC[4] and found 8,091 genomes with an annotated 'isolation source'. To identify laboratory strains, we searched for genomes with an isolation source annotated as laboratory, common laboratory strain or laboratory strain. We initially found 14 laboratory strains. However, on manually examining these genomes, we determined that one strain, *E. coli* ECO3347, was a uropathogenic strain[17].

To determine whether a genome contained an active or inactive *efeU*, we searched for the *efeU* gene product in PATRIC, along with its nucleotide length. We specified a genome as having an active *efeU* if only one *efeU* gene product was present and its length was 831 bp or longer. We specified a genome as having inactive *efeU* if two genes were annotated as *efeU*, and each gene product was shorter than 831 bp. These gene lengths were determined using *E. coli* strains

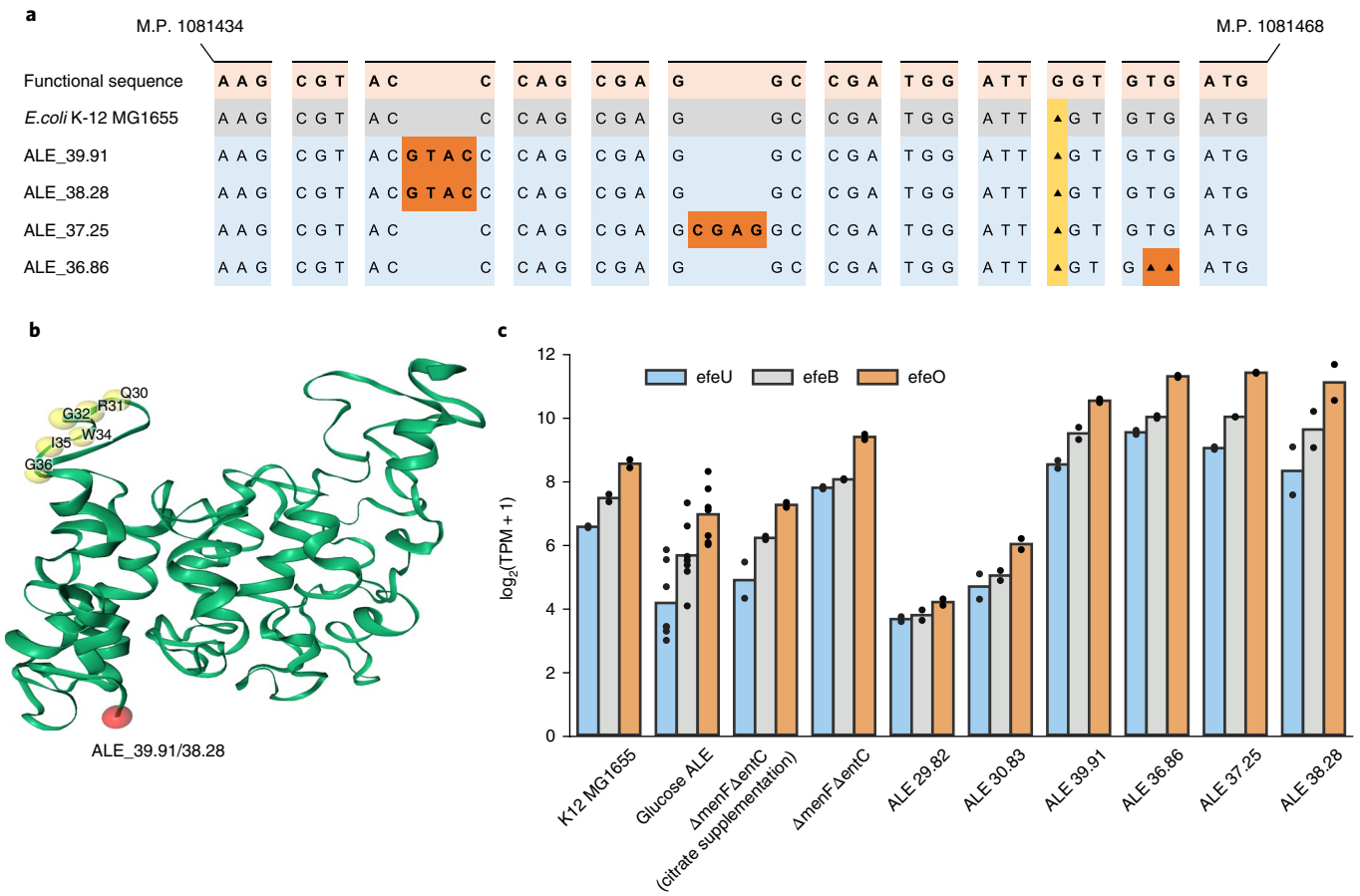

Fig. 2 | Mutations in *efeU* acquired during ALE. a, Mutations restoring the ORF of the *efeU* gene in independently evolved Δ*menF*Δ*entC* (ALE_39.91) and Δ*menF*Δ*entC*Δ*ubiC* (ALE_38.28, ALE_37.25 and ALE_36.86) strains. Map position (M.P.) corresponds to the *E. coli* K-12 MG1655 genome. Yellow and orange regions represent the positions of pseudogenic mutations and gene repairing mutations, respectively; triangles represent base deletions at the corresponding locations. **b**, Overlay of restored protein sequences on the homology model of functional EfeU. Black indicates deletion; yellow single-nucleotide polymorphism; green identical sequence; and red amino acid residue number 1. **c**, Expression profile of *efeUOB* operon in *E. coli* K-12 MG1655, glucose-adapted *E. coli* along with pre-evolved Δ*menF*Δ*entC*, evolved Δ*menF*Δ*entC* and Δ*menF*Δ*entC*Δ*ubiC* strains. Bars represent averages across replicates ($n = 7$ independent evolutionary end points for glucose ALE, and two biologically independent replicates for all others). TPM, transcripts per million.

MG1655 (fragmented), Nissle 1917 (active) and O157:H7 (active) as benchmarks[3]. Finally, we ignored the 20 genomes that had three or four genes annotated as *efeU*.

Patterns of efeU fragmentation. We downloaded the nucleotide sequences for all gene products ('genome features') in PATRIC with EfeU as the annotated product or gene symbol. We assumed all sequences with a nucleotide length of less than 831 bp to be potentially fragmented. We thus found 376 unique genome IDs that had a potentially fragmented *efeU* gene. Of these genomes, we kept the 234 that had exactly two annotated *efeU* gene products for further analysis. For each of these genomes, we computed the nucleotide lengths of each of the two fragments. The shorter of the two fragments had lengths ranging from 99 to 426 bp, with a clear mode at 120 bp (161 of 234 = 68.8%). The longer of the two fragments similarly had a clear mode at 720 bp (163 of 234 = 69.7%), with lengths ranging from 150 to 828 bp. Finally, we confirmed that all genomes annotated as MG1655 in the PATRIC genome had two *efeU* gene products with lengths of 120 and 720 bp, corresponding to the most frequent *efeU* fragment lengths observed across all *E. coli* genomes in the PATRIC database.

Enrichment test. To test for enrichment of the inactive *efeU* in laboratory strains, we used a one-sided Fisher's exact test using scipy.stats.fisher_exact in Python. From this test, we computed an odds ratio of 276 and $P = 1.8 \times 10^{-17}$. The one-sided 95% confidence interval for the odds ratio is (71.0, +Inf), which was computed using the exact2x2 R package[18].

Phylogeny test. The *rpoB* sequences of all the *E. coli* genomes used in this study were derived from the PATRIC database. We could obtain *rpoB* sequences of 212 genomes out of 234 genomes with the *efeU* gene fragmented into two parts. The multiple sequence alignment of *rpoB* sequences was performed using a program based on fast Fourier transform (MAFFT version 7)[19]. The *Escherichia fergusonii* (genome ID 585054.5) *rpoB* sequence was used as an outgroup for rooting the tree[20]. The evolutionary analyses were conducted using the same platform and the neighbour-joining method with bootstrap on (100 iterations to estimate recoverability of the nodes). The phylogenetic trees were visualized using FigTree version 1.4.3[21]. We isolated 1,187 unique *rpoB* sequences (4,029 bp) from the genomes with complete *efeU* genes. The whole genome tree was used to extract the genomes with a complete *efeU* gene clustering near genomes with a fragmented (120 bp and 720 bp) *efeU* gene. These extracted genomes along with the genomes with *efeU* fragmented into two parts were finally used to generate a more resolved tree.

Natural compensatory frameshift search. We examined the *efeU* sequence in 11,792 *E. coli* PATRIC genome IDs with one annotated *efeU* gene. We identified 11,620 (~98.5%) *E. coli* genome IDs with an *efeU* gene 831 bp in length and observed 598 unique *efeU* sequences. 102 genome IDs had an *efeU* sequence of 834,840or 846 bp. We performed sequence alignment on these using a multiple sequence alignment program based on fast Fourier transform (MAFFT version 7)[19].

ALE. ALE was performed using six independent replicates each of Δ*menF*Δ*entC* and Δ*menF*Δ*entC*Δ*ubiC*. Two independent replicates of Δ*menF*Δ*entC* were also used for evolution with iron supplementation. Cultures were serially propagated (150 µl passage volume) in 15 ml (working volume) flasks of M9 minimal medium with 4 g l^{-1} glucose, kept at 37 °C and well-mixed for full aeration. Iron supplementation during laboratory evolution was performed using 20 µM $FeSO_4$ and 10 mM sodium citrate[22]. An automated system passed the cultures to fresh flasks once they had reached an absorbance at 600 nm (A_{600}) of 0.3 (Tecan Sunrise plate reader, equivalent to an $A_{600} \approx 1$ on a traditional spectrophotometer with a 1 cm path length), a point at which nutrients were still in excess and exponential

growth had not started to taper off. Four A_{600} measurements were taken from each flask, and the slope of $\ln(A_{600})$ versus time determined the culture growth rates. A cubic interpolating spline constrained to be monotonically increasing was fitted to these growth rates to obtain the fitness trajectory curves.

DNA resequencing. DNA resequencing was performed on a clone from the end points of evolved strains. Total DNA was sampled from an overnight culture (1 ml of cell broth at an $A_{600} \approx 2.0$) and immediately centrifuged for 5 min at 8,000r.p.m. The supernatant was decanted, and the cell pellet was frozen at −80 °C. Genomic DNA was isolated using a Nucleospin Tissue kit (Macherey Nagel 740952.50) following the manufacturer's protocol, including treatment with RNase A. Resequencing libraries were prepared using a Nextera XT kit (Illumina FC-131–1024) following the manufacturer's protocol. Libraries were run on a HiSeq and/or NextSeq (Illumina).

Sequencing reads were filtered and trimmed using the software AfterQC version 0.9.6[23]. The breseq bioinformatics pipeline[24] version 0.31.1 was used to map sequencing reads and identify mutations relative to an *E. coli* K-12 MG1655 reference genome (NC_000913.3) amended to best reflect the starting strain[25]. Mutation analysis was performed using ALEdb[26].

Transcriptomics. Total RNA was sampled from duplicate cultures. All the strains were grown in M9 medium containing ferric chloride as an iron source except *ΔmenFΔentC*, which was grown with ferric chloride and sodium citrate. To obtain the expression profile of the *ΔmenFΔentC* strain in the citrate-deprived condition, we started the culture with supplemented media. At $A_{600} \approx 0.4$, we washed the cells with unsupplemented media and allowed them to grow without citrate until they reached an $A_{600} \approx 0.6$. Then, 3 ml of cell broth (at an $A_{600} \approx 0.6$) was immediately added to two volumes of Qiagen RNA-protect Bacteria Reagent (6 ml), vortexed for 5 s, incubated at room temperature for 5 min and immediately centrifuged for 10 min at 17,500 r.p.m. The supernatant was decanted, and the cell pellet was stored at −80 °C. Cell pellets were thawed and incubated with Ready-Lyse Lysozyme, SuperaseIn, protease K and 20% SDS for 20 min at 37 °C. Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit columns and following vendor procedures. An on-column DNase treatment was performed for 30 min at room temperature. RNA was quantified using a Nano drop and quality assessed by running an RNA nano chip on a bioanalyser. The ribosomal RNA was removed using an Illumina Ribo-Zero rRNA removal kit for Gram-negative bacteria. A KAPA Stranded RNA-Seq Kit (Kapa Biosystems KK8401) was used following the manufacturer's protocol to create sequencing libraries with an average insert length of around 300 bp. Libraries were run on a HiSeq and/or NextSeq (Illumina).

Expression profiling was performed as described in Seo et al.[9] Raw sequencing reads were mapped to the reference genome (NC_000913.3) using bowtie version 1.1.2[27] with a maximum insert size of 1,000 and a maximum of two maximum mismatches after trimming 3 bp at the 3' ends. Transcript abundance was quantified using summarizeOverlaps from the R GenomicAlignments package, with strand inversion for the deoxyuridine triphosphate (dUTP) protocol and strict intersection mode[28]. We then estimated the dispersion of each gene using DESeq2[29]. Transcripts per million were calculated by DESeq2. The final expression values were log-transformed $\log_2[\mathrm{TPM}+1]$ for visualization and analysis. Expression data for the wild-type and glucose ALE were accessed from GSE65643 and GSE61327, respectively, and were analysed using the same pipeline.

Structural analysis. The functional protein sequence of EfeU from *E. coli* Nissle 1917 was uploaded to I-TASSER to obtain a homology model structure[30–32]. I-TASSER model 1 was used as representative for the analysis. The sequences of other peptides/proteins were aligned to the representative sequence to analyse the single-nucleotide polymorphisms, insertions or deletions. Visualization was done using following colour code:

(1) Black: deletion in the given sequence when compared to the base model
(2) Red: amino acid residue number 1
(3) Yellow: single-nucleotide polymorphism (usually labelled)
(4) Green: part of the protein that is identical to the base model

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability

All code used to analyse the data is available at https://github.com/SBRG/mutation_analysis.

Data availability

DNA sequencing data from this study are available from the Sequence Read Archive database (SRA accession PRJNA505542). RNA sequencing data from this study are available from the Gene Expression Omnibus database under the accession number GSE122779.

Received: 6 July 2018; Accepted: 5 December 2018;
Published online: 28 January 2019

References

1. Ratledge, C. & Dover, L. G. *Annu. Rev. Microbiol.* **54**, 881–941 (2000).
2. Pi, H. & Helmann, J. D. *Proc. Natl Acad. Sci. USA* **114**, 12785–12790 (2017).
3. Grosse, C. et al. *Mol. Microbiol.* **62**, 120–131 (2006).
4. Wattam, A. R. et al. *Nucleic Acids Res.* **45**, D535–D542 (2017).
5. Conrad, T. M., Lewis, N. E. & Palsson, B. Ø. *Mol. Syst. Biol.* **7**, 509 (2011).
6. Sandberg, T. E. et al. *PLoS ONE* **11**, e0151130 (2016).
7. Zhou, K., Aertsen, A. & Michiels, C. W. *FEMS Microbiol. Rev.* **38**, 119–141 (2014).
8. Miethke, M., Monteferrante, C. G., Marahiel, M. A. & van Dijl, J. M. *Biochim. Biophys. Acta* **1833**, 2267–2278 (2013).
9. Seo, S. W. et al. *Nat. Commun.* **5**, 4910 (2014).
10. Tutar, Y. *Comp. Funct. Genomics* **2012**, 424526 (2012).
11. Kuo, C. H. & Ochman, H. *PLoS Genet.* **6**, e1001050 (2010).
12. Lawrence, J. G., Hendrix, R. W. & Casjens, S. *Trends Microbiol.* **9**, 535–540 (2001).
13. Mira, A., Ochman, H. & Moran, N. A. *Trends Genet.* **17**, 589–596 (2001).
14. Thomason, L. C., Costantino, N. & Court, D. L. *Curr. Protoc. Mol. Biol.* **79**, 1.17.1–1.17.8 (2007).
15. Baba, T. et al. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
16. Datsenko, K. A. & Wanner, B. L. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
17. Zhou, K. et al. *Int. J. Antimicrob. Agents* **51**, 822–828 (2018).
18. Fay, M. P. *Biostatistics* **11**, 373–374 (2010).
19. Katoh, K., Rozewicki, J. & Yamada, K. D. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbx108 (2017).
20. Touchon, M. et al. *PLoS Genet.* **5**, e1000344 (2009).
21. Rambaut, A. FigTree v1.4.3 (Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, 2018); http://tree.bio.ed.ac.uk/software/figtree/
22. Fleming, T. P., Nahlik, M. S. & McIntosh, M. A. *J. Bacteriol.* **156**, 1171–1177 (1983).
23. Chen, S. et al. *BMC Bioinform.* **18**, 80 (2017).
24. Deatherage, D. E. & Barrick, J. E. *Methods Mol. Biol.* **1151**, 165–188 (2014).
25. Phaneuf, P. Zenodo v.1.4.1 (Zenodo, San Diego, 2018); https://doi.org/10.5281/zenodo.1301237
26. Phaneuf, P. V., Gosting, D., Palsson, B. & Feist, A. Preprint at *bioRxiv* (2018); https://www.biorxiv.org/content/biorxiv/early/2018/05/15/320747.full.pdf
27. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. *Genome. Biol.* **10**, R25 (2009).
28. Lawrence, M. et al. *PLoS Comput. Biol.* **9**, e1003118 (2013).
29. Love, M. I., Huber, W. & Anders, S. *Genome Biol.* **15**, 550 (2014).
30. Yang, J. et al. *Nat. Methods* **12**, 7–8 (2015).
31. Roy, A., Kucukural, A. & Zhang, Y. *Nat. Protoc.* **5**, 725–738 (2010).
32. Zhang, Y. *BMC Bioinform.* **9**, 40 (2008).

Acknowledgements

This work was funded by the Novo Nordisk Foundation under grant number NNF10CC1016517.

Author contributions

A.A., A.M.F. and B.O.P. designed the study. A.A., C.A.O., S.X., Y.H. and R.S. performed the experiments. A.A., L.Y., A.V.S., C.A.O., E.C. and T.E.S. analysed the data. K.S.C. and P.V.P. contributed analysis tools. A.A. and B.O.P. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-018-0340-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to B.O.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Bernhard O. Palsson

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	All data analysis was performed using Python 2.7 using standard Numpy (v1.14.1) and Scipy (v1.0.0) modules. The ipython (jupyter) notebooks used to perform analysis is available online on a public github repository: https://github.com/SBRG/mutation_analysis. DNA re-sequencing analysis was performed using AfterQC version 0.9.6 and breseq bioinformatics pipeline version 0.31.1. Bowtie v1.1.2, summarizeOverlaps from the R GenomicAlignments package and DESeq2 were used for RNA-Seq analysis. Multiple sequence alignment and evolutionary analysis were performed using MAFFT version 7. The phylogenetic tree was visualized using FigTree v1.4.3

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data is available in the main text or the supplementary materials. DNA sequencing and RNA-Seq data accession details are provided.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample sizes. Laboratory evolution experiments were performed with six independent replicates which is a standard practice to effectively run the Adaptive Laboratory Evolution platform.
Data exclusions	No exclusion
Replication	Independent biological replicates were used and reproducible data were obtained for each replicate. One of the six independent replicates of ΔmenFΔentC strain and three out of the six replicates of ΔmenFΔentCΔubiC acquired pseudogene repair related mutation. This is because of the random nature of mutation.
Randomization	Experimental group allocation was not needed
Blinding	No such group allocation was required in this study

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging