Title

Revealing Causes for False-Positive and False-Negative Calling of Gene Essentiality in Escherichia coli Using Transposon Insertion Sequencing

Journal

mSystems, 8(1)

ISSN

2379-5077

Authors

Choe, Donghui

Kim, Uigi

Hwang, Soonkyu

et al.

Publication Date

2023-02-23

DOI

10.1128/msystems.00896-22

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

OBSERVATION

Revealing Causes for False-Positive and False-Negative Calling of Gene Essentiality in *Escherichia coli* Using Transposon Insertion Sequencing

Donghui Choe,[a] Uigi Kim,[b] Soonkyu Hwang,[b] Sang Woo Seo,[c] Donghyuk Kim,[d] Suhyung Cho,[b] Bernhard Palsson,[a,e] Byung-Kwan Cho[b]

[a]Department of Bioengineering, University of California San Diego, La Jolla, California, USA
[b]Department of Biological Sciences and KI for the BioCentury, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[c]School of Chemical and Biological Engineering, Seoul National University, Seoul, Republic of Korea
[d]School of Energy and Chemical Engineering, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea
[e]Department of Pediatrics, University of California San Diego, La Jolla, California, USA

ABSTRACT The massive sequencing of transposon insertion mutant libraries (Tn-Seq) represents a commonly used method to determine essential genes in bacteria. Using a hypersaturated transposon mutant library consisting of 400,096 unique Tn insertions, 523 genes were classified as essential in *Escherichia coli* K-12 MG1655. This provided a useful genome-wide gene essentiality landscape for rapidly identifying 233 of 301 essential genes previously validated by a knockout study. However, there was a discrepancy in essential gene sets determined by conventional gene deletion methods and Tn-Seq, although different Tn-Seq studies reported different extents of discrepancy. We have elucidated two causes of this discrepancy. First, 68 essential genes not detected by Tn-Seq contain nonessential subgenic domains that are tolerant to transposon insertion, which leads to the false assignment of an essential gene as a nonessential or dispensable gene. These genes exhibited a high level of transposon insertion in their subgenic nonessential domains. In contrast, 290 genes were additionally categorized as essential by Tn-Seq, although their knockout mutants were available. The comparative analysis of Tn-Seq and high-resolution footprinting of nucleoid-associated proteins (NAPs) revealed that a protein-DNA interaction hinders transposon insertion. We identified 213 false-positive genes caused by NAP-genome interactions. These two limitations have to be considered when addressing essential bacterial genes using Tn-Seq. Furthermore, a comparative analysis of high-resolution Tn-Seq with other data sets is required for a more accurate determination of essential genes in bacteria.

IMPORTANCE Transposon mutagenesis is an efficient way to explore gene essentiality of a bacterial genome. However, there was a discrepancy between the essential gene set determined by transposon mutagenesis and that determined using single-gene knockout strains. In this study, we generated a hypersaturated *Escherichia coli* transposon mutant library comprising approximately 400,000 different mutants. Determination of transposon insertion sites using next-generation sequencing provided a high-resolution essentiality landscape of the *E. coli* genome. We identified false negatives of essential gene discovery due to the permissive insertion of transposons in the C-terminal region. Comparisons between the transposon insertion landscape with binding profiles of DNA-binding proteins revealed interference of nucleoid-associated proteins to transposon insertion, generating false positives of essential gene discovery. Consideration of these findings is required to avoid the misinterpretation of transposon mutagenesis results.

KEYWORDS gene essentiality, subgenic-level essentiality, Tn-Seq, DNA-binding proteins, nucleoid-associated proteins

Editor Daniel Garrido, Pontificia Universidad Catolica de Chile

Address correspondence to Bernhard Palsson, bpalsson@ucsd.edu, or Byung-Kwan Cho, bcho@kaist.ac.kr.

The authors declare no conflict of interest.

Received 15 September 2022
Accepted 18 November 2022
Published 12 December 2022

Determination of the essential set of genes encoded in a bacterial genome is an important part of understanding the genotype-phenotype relationship, engineering their metabolic capabilities, and creating synthetic bacterial genomes (1, 2). Several methods have been devised to discriminate between the essential and dispensable genes in a bacterium (3–6). Among those methods, transposon mutagenesis coupled with next-generation sequencing (Tn-Seq) identifies dispensable genes based on the generation of disruptive genomic insertions in a high-throughput manner (7–9). However, essential genes determined by transposon mutagenesis (10) differ from those determined by conventional knockout studies (4, 11, 12). Although some studies have reported a low rate of discrepancy (13), it varies from study to study (10), indicating there may be an inherent limitation to the method. The discrepancy originates from different sources, such as domain essentiality, sequence preference of transposon, and nucleoid-associated protein (NAP) interference (14–17). Here, we have revisited the relationship between the lack of transposon insertion and the interaction of NAPs with the genome for the accurate determination of essential genes in *Escherichia coli* K-12 MG1655 using a hypersaturated transposon mutant library.

We constructed 1 million transposon insertion mutants of *E. coli* K-12 MG1655 capable of growing on solid LB medium (see Text S1 in the supplemental material). Unique transposon insertion sites (TISs) across its genome were determined by using massively parallel sequencing. From 2.4×10^6 mapped reads, a high-resolution transposon insertion landscape consisting of 400,096 unique TISs was obtained (Fig. 1A). TISs were distributed evenly across the genome, so that 89.6% had an adjacent TIS within 10 bp (Fig. 1B). Furthermore, GC content near TISs was 55.0%, which indicated that insertion bias based on GC content was negligible, considering that the GC content of the genome is 50.8%. Also, genomic regions with high or low GC content ($>$65% or $<$15%) showed no lack of transposon insertions, which are known to be depleted during amplification (Fig. S1 and Text S1) (18). As expected, we observed that essential genes, such as RNA polymerase subunits (*rpoBC*), had low transposon insertion frequencies (TIFs), whereas the nonessential genes had higher TIFs (Fig. 1C) (10, 19, 20).

To measure the TIF of each gene, we defined an insertion per kilobase per million insertions (IPKM) value, which is a modification of the widely used reads per kilobase per million mapped reads (RPKM) metric (Fig. S2A and B, Text S1) (21). The TISs that mapped onto either end of genes (4% at each end) were excluded from TIF calculation, since essential genes are tolerant to such insertions (Fig. 1D and Fig. S2C). To test the robustness of the end-curated IPKM (ecIPKM) to estimate gene essentiality (Table S1), a known set of essential genes in *E. coli* according to the Profiling of *Escherichia coli* Chromosome (PEC) database (11) and pseudogenes were examined (Fig. 1E). The ecIPKM of all genes ranged from 0 to 7,208.7 (median of 95.2), whereas PEC genes had very low ecIPKM values (median of 0.858). In contrast, pseudogenes exhibited high ecIPKM values (median of 52.1), in agreement with the fact that these genes are regarded as nonfunctional.

To define a threshold that determined essential genes, the accuracies of ecIPKM cutoffs discovering the PEC essential genes were examined. The end curation improved the accuracy of essential gene discovery (Fig. 1F) without affecting overall IPKM distribution (Fig. S2D and E). We defined ecIPKM of 2.2 as a cutoff where accuracy was maximized (Text S1, Fig. 1G, and Fig. S2F and G). Using this criterion, 523 genes were determined as essential, of which 233 were PEC essential genes (Fig. 1H and Table S1). Compared with a previous transposon mapping with lower mutant library density (10) and Tn-Seq (13), which reported 68.8% and 88.7% coverage of PEC genes with false discoveries (n = 406 and 92), respectively, we identified similar coverage (77.4%) and false discoveries (n = 290). Essential gene sets discovered by different approaches seemed to vary due to many factors, such as experimental procedures, conditional essentiality, and statistics of analysis (Text S1 and Fig. S3).

Tn-Seq failed to detect 68 essential genes. Tn-Seq with the ecIPKM metric failed to detect 68 PEC essential genes (Fig. 1H). These false negatives have been investigated previously (13). The comparison revealed that 22 genes were also falsely classified. This was caused by tolerated insertions on subgenic elements of essential genes, a polar effect, and misclassification in the PEC data set, among other things. Manual

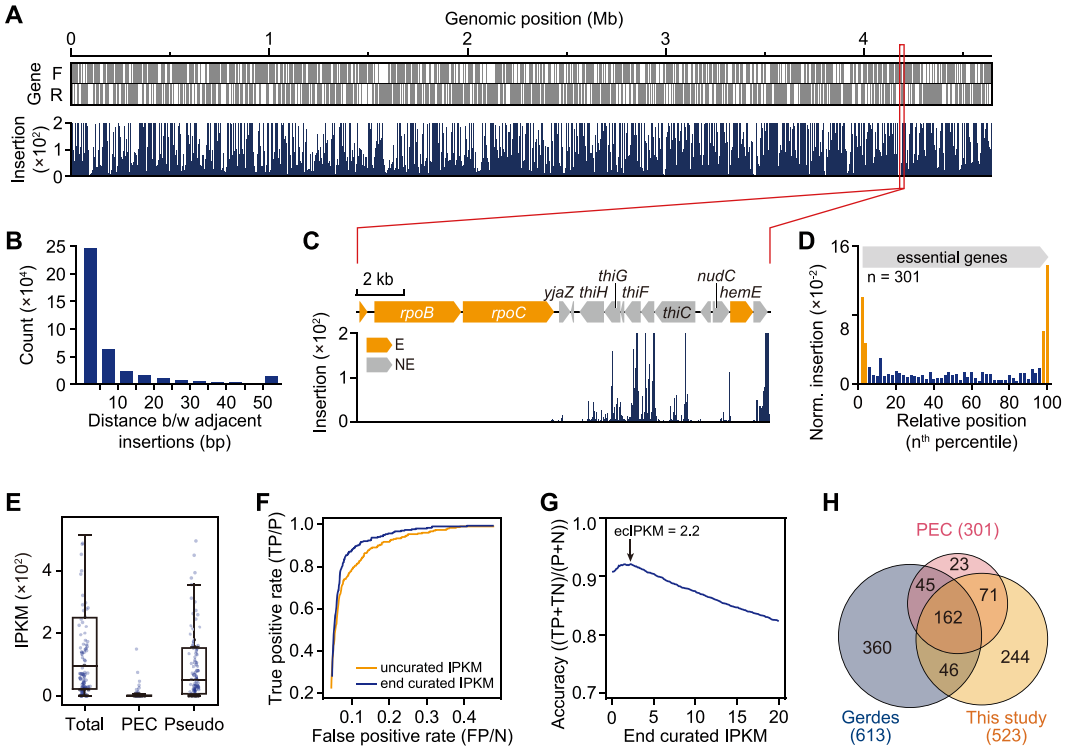

FIG 1 Genome-wide determination of gene essentiality using Tn-Seq. (A) Transposon insertion landscape of the *E. coli* MG1655 genome. Insertion profiles shown in this figure indicate normalized insertion (insertions per million insertions). (B) Histogram showing distances between two adjacent insertions. Over 90% of the insertion had an adjacent insertion within 25 bp. (C) Essential genes were not tolerant to transposon insertion due to disruptive properties of the transposon. E and NE indicate essential and nonessential genes, respectively, as reported in the PEC database. (D) Transposon insertion profile of PEC essential genes. (E) IPKM distribution of the total 4,498 genes, PEC essential genes, and pseudogenes. (F) A receiver operating characteristic (ROC) curve of IPKM when discovering essential genes. (G) Accuracy of essential gene discovery with different ecIPKM cutoffs. (H) Comparison between PEC essential genes and essential genes determined by previous Tn-Seq (Gerdes) and by this study, showing considerable discrepancies between experiments. Six phantom genes (b0322, b1228, b2612, b2651, b3112, and b4229) were excluded from the analysis.

inspection of the false negatives unique to this study revealed that the false classifications were due to the same causes (Text S1 and Table S2). Most of the false negatives were caused by insertions on nonessential domains of essential genes, as recapitulated in a complementation experiment (Fig. S4).

DNA-binding proteins interfered with transposon insertion. Next, we analyzed 290 false essential genes determined in the Tn-Seq library constructed from LB medium whose deletion mutants have been reported previously (Fig. 1H) (4, 11, 12). For example, a genomic region containing nonessential genes was protected from transposon insertion (Fig. 2A). We hypothesized that DNA-binding proteins (DBPs) can interfere with transposon insertion, as shown previously in different bacteria (16, 17). Specifically, we focused on NAP-DNA interactions, which were maintained similarly throughout the growth phase (22). Genome-wide binding regions of six NAPs in *E. coli* grown in M9 glucose medium were obtained by chromatin immunoprecipitation with exonuclease digestion (ChIP-Exo) (23). A total of 3,669 NAP-binding regions were detected ($q < 0.05$, Model-based Analysis of ChIP-Seq (MACS2) software) (Table S3) and compared with Tn-Seq data obtained from the same medium (Table S1). The NAP-binding regions perfectly overlapped with the protected regions from transposon insertion (Fig. 2B to F). Statistical analysis indicated the ecIPKM values of the NAP-binding regions were markedly lower ($P < 0.05$, Welch's *t* test) than those for the random genomic regions (Fig. 2G). This was not due to specific bindings of NAPs to essential genes, because H-NS and StpA bound only 11 and 13 PEC genes, respectively, whereas the randomly sampled genomic regions overlapped 54.7 and 67.3 PEC genes on average, respectively. Overall, most of the false positives (238/290; 82.1%) contained NAP-binding regions covering more

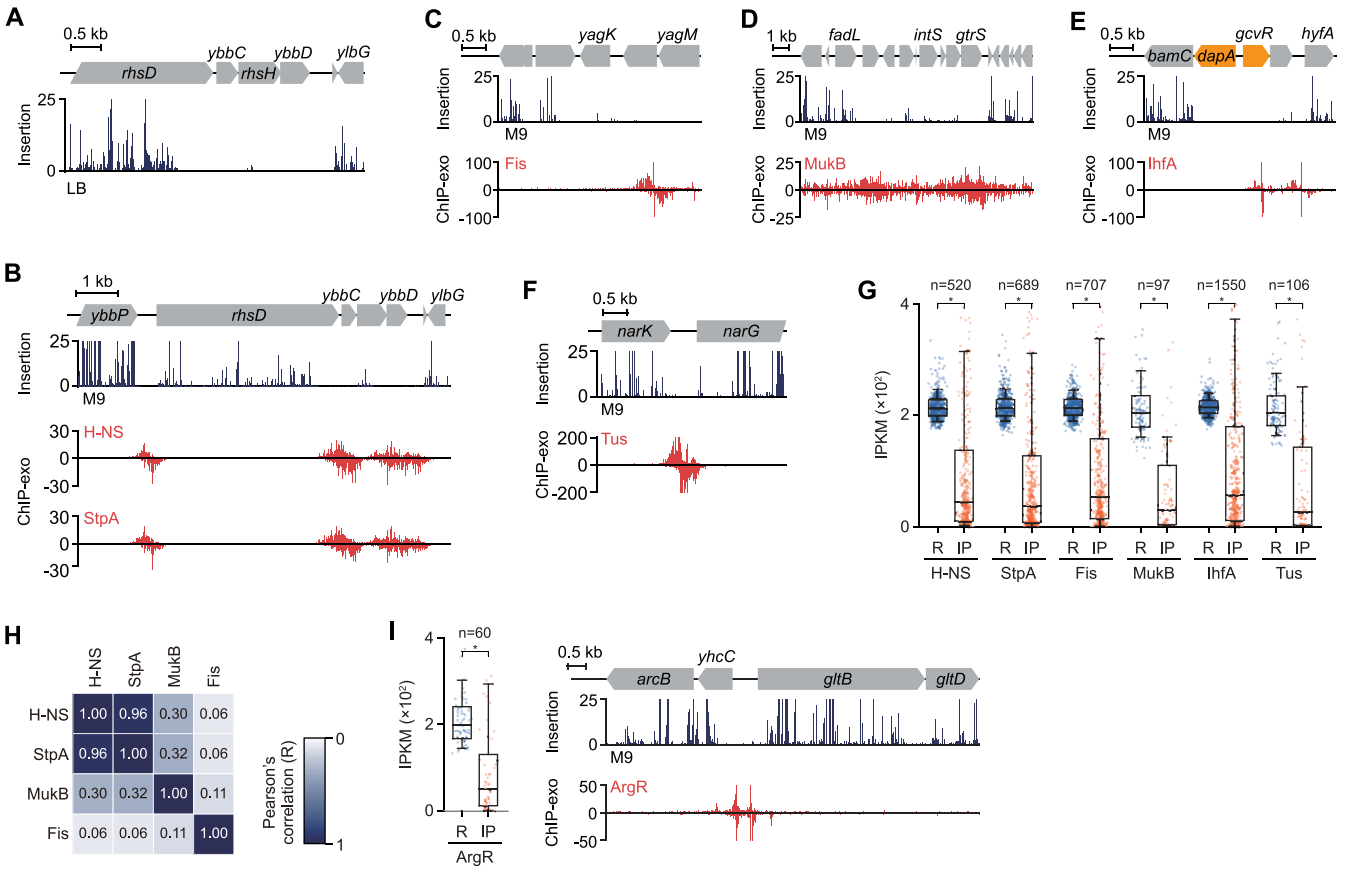

FIG 2 Interference of DNA-binding protein with transposon insertion revealed by ChIP-Exo. (A) Transposon insertion was not observed in the genomic region containing three nonessential genes, *ybbC*, *rhsH*, and *ybbD*. The profile shows Tn-Seq in LB medium. (B) Profiles show transposon insertion and binding of H-NS and StpA for three nonessential genes, *ybbC*, *rhsH*, and *ybbD*. H-NS and StpA binding on the three genes overlapped with the transposon-protected region. (C to F) Transposon insertion and ChIP-Exo profiles of Fis (C), MukB (D), IhfA (E), and Tus (F). (G) Box plot showing IPKM distribution of NAP-binding regions. Dots indicate IPKM of each region. As a control, the same-sized random genomic regions were tested. One of the 10,000 bootstrapped set is shown. R, random genomic regions; IP, NAP-binding regions determined by ChIP-Exo. Statistical significance was assessed by comparing the bootstrapped distribution of mean IPKM of 10,000 random sets. *, $P < 0.05$ (Welch's *t* test). (H) Heatmap showing pairwise comparisons between H-NS, StpA, MukB, and Fis profiles. Pearson's correlation coefficient (*R*) was calculated from ChIP-Exo intensity across the genome. Bindings of H-NS and StpA correlated with a coefficient of 0.955. MukB showed a relatively high correlation (>0.3) with H-NS and StpA, compared to Fis. (I) Box plot showing IPKM distribution of ArgR-binding regions; profiles show transposon insertion and ArgR binding near *gltB*. See the legend for panel G for details.

than 80% of the genic region, although the two data sets were collected from cells grown in different media. If we assumed the NAP binding was independent of medium, this would indicate that the interference is a major cause of false-positive discovery. The remaining 52 false positives were likely a result of DBPs that were not examined in this study that interfered with transposon insertion.

Although we observed a few positions that were deprotected for transposon insertion in a Δ*hns* Δ*stpA* double mutant (Fig. S5A and B and Table S4), transposon insertion of the NAP-binding regions was not fully relieved (Fig. S5C and D and Text S1), unlike the previous report on different bacteria (16). Even when the two NAPs were deleted simultaneously, other NAPs may have complemented the function of the two. MukB could be a candidate, as its binding profile is comparable to those of H-NS and StpA ($R > 0.3$) (Fig. 2H and Fig. S5E). In addition to the NAPs, the DNA-structuring transcriptional regulator ArgR (24) showed the same effect (Fig. 2I), which indicated that transcription factors also participated in this interference. We concluded that the interaction between DBPs and the bacterial genome interferes with transposon insertion, partially explaining the inconsistency between Tn-Seq and the knockout study when assessing gene essentiality. Overall, reevaluation of Tn-Seq results revealed that a consideration of interference of transposon insertion by DBPs is required to avoid misinterpretation of results. Unfortunately, there is no *E. coli* strain that lacks all the NAPs, nor

is there a way of preventing DNA-NAP interactions to rapidly screen false essential genes. Thus, a high-density Tn-Seq experiment and careful evaluation of the results are necessary. Our study not only revealed the complexity of transposon insertion that leads to the identification of false essential genes but also has provided a high-resolution gene essentiality landscape of the *E. coli* genome.

Data availability. The sequencing data have been deposited in the European Nucleotide Archive (accession number PRJEB22130).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, DOCX file, 0.1 MB.
FIG S1, JPG file, 1 MB.
FIG S2, JPG file, 2.4 MB.
FIG S3, JPG file, 1.1 MB.
FIG S4, JPG file, 2.7 MB.
FIG S5, JPG file, 2 MB.
TABLE S1, XLSX file, 0.7 MB.
TABLE S2, XLSX file, 0.02 MB.
TABLE S3, XLSX file, 0.1 MB.
TABLE S4, XLSX file, 0.04 MB.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation (grants 2021M3A9I4024308 and 2021M3A9I5023245 to B.-K.C. and 2021R1A2C1012589 to S.C.) funded by the Ministry of Science and ICT, Republic of Korea, and the Novo Nordisk Foundation (NNF) Center for Biosustainability (grant NNF16CC0021858 to B.P.) at the Technical University of Denmark. We thank Marc Abrams for editing the manuscript.

B.P. and B.-K.C. designed and supervised the project; D.C., U.K., S.H., S.W.S., D.K., and S.C. performed experiments; D.C., S.C., B.P., and B.-K.C. analyzed the data; D.C., S.C., B.P., and B.-K.C. wrote the manuscript. All authors have read and approved the final manuscript.

We declare no competing interests.

REFERENCES

1. Cho BK, Palsson BO. 2009. Probing the basis for genotype-phenotype relationships. Nat Methods 6:565–566. https://doi.org/10.1038/nmeth0809-565.
2. Hutchison CA, III, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L, Pelletier JF, Qi ZQ, Richter RA, Strychalski EA, Sun L, Suzuki Y, Tsvetanova B, Wise KS, Smith HO, Glass JI, Merryman C, Gibson DG, Venter JC. 2016. Design and synthesis of a minimal bacterial genome. Science 351:aad6253. https://doi.org/10.1126/science.aad6253.
3. Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC. 1999. Global transposon mutagenesis and a minimal *Mycoplasma* genome. Science 286:2165–2169. https://doi.org/10.1126/science.286.5447.2165.
4. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio Collection. Mol Syst Biol 2:2006.0008. https://doi.org/10.1038/msb4100050.
5. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BO, Agarwalla S. 2006. Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. J Bacteriol 188:8259–8271. https://doi.org/10.1128/JB.00740-06.
6. Rousset F, Cui L, Siouve E, Becavin C, Depardieu F, Bikard D. 2018. Genome-wide CRISPR-dCas9 screens in *E. coli* identify essential genes and phage host factors. PLoS Genet 14:e1007749. https://doi.org/10.1371/journal.pgen.1007749.
7. Gallagher LA, Shendure J, Manoil C. 2011. Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. mBio 2:e00315-10. https://doi.org/10.1128/mBio.00315-10.
8. Ibberson CB, Stacy A, Fleming D, Dees JL, Rumbaugh K, Gilmore MS, Whiteley M. 2017. Co-infecting microorganisms dramatically alter pathogen gene essentiality during polymicrobial infection. Nat Microbiol 2:17079. https://doi.org/10.1038/nmicrobiol.2017.79.
9. van Opijnen T, Bodi KL, Camilli A. 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods 6:767–772. https://doi.org/10.1038/nmeth.1377.
10. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, Osterman AL. 2003. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. J Bacteriol 185:5673–5684. https://doi.org/10.1128/JB.185.19.5673-5684.2003.
11. Hashimoto M, Ichimura T, Mizoguchi H, Tanaka K, Fujimitsu K, Keyamura K, Ote T, Yamakawa T, Yamazaki Y, Mori H, Katayama T, Kato J. 2005. Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. Mol Microbiol 55:137–149. https://doi.org/10.1111/j.1365-2958.2004.04386.x.
12. Kato J, Hashimoto M. 2007. Construction of consecutive deletions of the *Escherichia coli* chromosome. Mol Syst Biol 3:132. https://doi.org/10.1038/msb4100174.
13. Goodall ECA, Robinson A, Johnston IG, Jabbari S, Turner KA, Cunningham AF, Lund PA, Cole JA, Henderson IR. 2018. The essential genome of *Escherichia coli* K-12. mBio 9:e02096-17. https://doi.org/10.1128/mBio.02096-17.
14. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. 2012. Insertion site preference of Mu, Tn5, and Tn7 transposons. Mob DNA 3:3. https://doi.org/10.1186/1759-8753-3-3.

15. Lidholm DA, Lohe AR, Hartl DL. 1993. The transposable element mariner mediates germline transformation in *Drosophila melanogaster*. Genetics 134:859–868. https://doi.org/10.1093/genetics/134.3.859.
16. Kimura S, Hubbard TP, Davis BM, Waldor MK. 2016. The nucleoid binding protein H-NS biases genome-wide transposon insertion landscapes. mBio 7:e01351-16. https://doi.org/10.1128/mBio.01351-16.
17. Manna D, Porwollik S, McClelland M, Tan R, Higgins NP. 2007. Microarray analysis of Mu transposition in *Salmonella enterica*, serovar Typhimurium: transposon exclusion by high-density DNA binding proteins. Mol Microbiol 66:315–328. https://doi.org/10.1111/j.1365-2958.2007.05915.x.
18. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12:R18. https://doi.org/10.1186/gb-2011-12-2-r18.
19. van Opijnen T, Camilli A. 2012. A fine scale phenotype-genotype virulence map of a bacterial pathogen. Genome Res 22:2541–2551. https://doi.org/10.1101/gr.137430.112.
20. Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Coller JA, Fero MJ, McAdams HH, Shapiro L. 2011. The essential genome of a bacterium. Mol Syst Biol 7:528. https://doi.org/10.1038/msb.2011.58.
21. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628. https://doi.org/10.1038/nmeth.1226.
22. Kahramanoglou C, Seshasayee AS, Prieto AI, Ibberson D, Schmidt S, Zimmermann J, Benes V, Fraser GM, Luscombe NM. 2011. Direct and indirect effects of H-NS and Fis on global gene expression control in *Escherichia coli*. Nucleic Acids Res 39:2073–2091. https://doi.org/10.1093/nar/gkq934.
23. Seo SW, Gao Y, Kim D, Szubin R, Yang J, Cho BK, Palsson BO. 2017. Revealing genome-scale transcriptional regulatory landscape of OmpR highlights its expanded regulatory roles under osmotic stress in *Escherichia coli* K-12 MG1655. Sci Rep 7:2181. https://doi.org/10.1038/s41598-017-02110-7.
24. Cho S, Cho YB, Kang TJ, Kim SC, Palsson B, Cho BK. 2015. The architecture of ArgR-DNA complexes at the genome-scale in *Escherichia coli*. Nucleic Acids Res 43:3079–3088. https://doi.org/10.1093/nar/gkv150.