


Testing the adaptive hypothesis of lagging-strand encoding in bacterial genomes

Haoxuan Liu¹ & Jianzhi Zhang ¹✉

ARISING FROM Christopher N. Merrikh & Houra Merrikh. *Nature Communications* <https://doi.org/10.1038/s41467-018-07110-3> (2018)

Genes are preferentially encoded on the leading instead of the lagging strand of DNA replication in most bacterial genomes¹, likely because lagging-strand encoding is selectively disfavored. Merrikh and Merrikh, however, proposed that lagging-strand encoding is adaptive, based on their inferred gene inversions and a comparison of nucleotide substitution rates². Here we point out methodological flaws and errors in their analyses and logical problems of their interpretation. Our new analysis of their data and analysis of other publicly available data do not support the adaptive hypothesis of lagging-strand encoding.

Lagging-strand encoding can cause head-on (HO) collisions between DNA polymerases and RNA polymerases that induce transcriptional abortion, replication delay, and possibly mutagenesis¹, so is expected to be deleterious relative to leading-strand encoding. However, there are still genes encoded on the lagging strand, an observation that has been explained by a balance between deleterious mutations bringing genes from the leading to the lagging strand and purifying selection purging such mutations^{3,4}. This mutation-selection balance hypothesis predicts that the probability that a gene is encoded on the lagging strand decreases with the detriment of its lagging-strand encoding relative to leading-strand encoding, explaining why highly expressed genes and essential genes are underrepresented on the lagging strand^{5,6}. By contrast, Merrikh and Merrikh² asserted that the observed lagging-strand encoding is adaptive because of beneficial mutations brought by the potentially increased mutagenesis resulting from HO collisions.

Following Merrikh and Merrikh², we refer to the leading-strand encoded genes as co-directional (CD) genes, because the movement of DNA and RNA polymerases in these genes is CD, and refer to lagging-strand encoded genes as HO genes. Merrikh and Merrikh's primary evidence for the adaptive hypothesis was their inference that the fraction of present-day HO genes that were previously CD exceeds the fraction of present-day CD genes that were previously HO in each of the six bacterial species examined, which they interpreted as natural selection for HO under the reasonable assumption that the inversion mutation from HO to CD and that from CD to HO have equal rates per

gene. However, if this interpretation were true, the number of HO genes would gradually rise and eventually exceed the number of CD genes, contradicting the preponderance of CD genes in most bacterial genomes. The cause of the above paradox is that to infer selection, one should compare the CD-to-HO inversion rate with the HO-to-CD inversion rate. But the CD-to-HO rate does not equal the fraction of present-day HO genes that were CD, but the fraction of previously CD genes that are now HO. The same can be said for the HO-to-CD rate. Therefore, the fractions estimated by Merrikh and Merrikh have no bearing on the hypothesis being tested. Based on Merrikh and Merrikh's estimates of previously and present-day HO genes and CD genes (data from Supplementary Table 1 in Merrikh and Merrikh²), we found that the rate of inversion from CD to HO is lower than the reverse rate in four of the six species examined (Table 1), consistent with the prediction of the mutation-selection balance hypothesis but opposite to that of the adaptive hypothesis. In fact, a lower inversion rate from CD to HO than the converse rate was reported 20 years ago in each of the four bacterial species examined then⁷. Merrikh and Merrikh's mistake is puzzling given that they cited this study in their paper².

Furthermore, Merrikh and Merrikh's inference of previously HO genes and previously CD genes² is error-prone. For each genic region, they computed the leading-strand GC skew by the difference in frequency between guanine (G) and cytosine (C), relative to the total frequency of G and C. They assumed that a negative GC skew means that the gene has been recently inverted², based on a tendency for the leading strand to have a positive GC skew due to mutational bias⁸. It should be noted that the G and C frequencies of a genic region are influenced not only by mutation but also by selection⁹. To the best of our knowledge, no study has used the GC skew alone to infer gene inversion in bacteria. Comparison of gene orientation determined from genome assemblies is the standard method to infer gene inversion in bacterial (and organelle) genomes, while the GC skew is sometimes used as confirmatory evidence^{7,10,11}. Even when the GC skew is used as a confirmation, the common practice is to calculate it at third codon positions or four-fold degenerate sites, because nucleotide frequencies at these sites are subject to weaker

¹Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, United States. ✉email: jianzhi@umich.edu

Table 1 HO-to-CD and CD-to-HO inversion rates based on the inversions inferred by Merrikh and Merrikh.

| Species | Current no. of CD genes | Current no. of HO genes | HO-to-CD inversions | CD-to-HO inversions | HO-to-CD inversion rate ^a | CD-to-HO inversion rate ^b | P value ^c |
|-----------------------------------|-------------------------|-------------------------|---------------------|---------------------|--------------------------------------|--------------------------------------|------------------------|
| <i>Mycoplasma gallisepticum</i> | 651 | 159 | 166 | 121 | 0.81 | 0.20 | <10 ⁻⁵ |
| <i>Staphylococcus aureus</i> | 2216 | 656 | 84 | 428 | 0.27 | 0.17 | <10 ⁻⁵ |
| <i>Bacillus subtilis</i> | 3104 | 1139 | 127 | 777 | 0.26 | 0.21 | 7.4 × 10 ⁻³ |
| <i>Campylobacter jejuni</i> | 1048 | 651 | 23 | 529 | 0.16 | 0.34 | <10 ⁻⁵ |
| <i>Klebsiella pneumoniae</i> | 2999 | 2405 | 465 | 1102 | 0.26 | 0.30 | 2.3 × 10 ⁻³ |
| <i>Mycobacterium tuberculosis</i> | 2415 | 1665 | 603 | 712 | 0.39 | 0.28 | <10 ⁻⁵ |

^aNumber of HO-to-CD inversions divided by the number of previously HO genes, which is the number of current HO genes plus the number of HO-to-CD inversions minus the number of CD-to-HO inversions.

^bNumber of CD-to-HO inversions divided by the number of previously CD genes, which is the number of current CD genes plus the number of CD-to-HO inversions minus the number of HO-to-CD inversions.

^cChi-squared test of equal rates of HO-to-CD and CD-to-HO inversions, performed using a 2 × 2 table of the number of previously HO genes that remain HO, number of HO-to-CD inversions, number of previously CD genes that remain CD, and the number of CD-to-HO inversions. The P values have not been corrected for multiple testing.

Table 2 Species used to infer gene inversions.

| Groups | Species of interest (A) | Sister species (B) | Outgroup species (C) |
|--------|-----------------------------------|-------------------------------------|----------------------------------|
| 1 | <i>Mycoplasma gallisepticum</i> | <i>Mycoplasma genitalium</i> | <i>Mycoplasma penetrans</i> |
| 2 | <i>Staphylococcus aureus</i> | <i>Staphylococcus saprophyticus</i> | <i>Staphylococcus sciuri</i> |
| 3 | <i>Bacillus subtilis</i> | <i>Bacillus licheniformis</i> | <i>Bacillus cereus</i> |
| 4 | <i>Campylobacter jejuni</i> | <i>Campylobacter lari</i> | <i>Campylobacter ureolyticus</i> |
| 5 | <i>Klebsiella pneumoniae</i> | <i>Klebsiella oxytoca</i> | <i>Escherichia coli</i> |
| 6 | <i>Mycobacterium tuberculosis</i> | <i>Mycobacterium haemophilum</i> | <i>Mycobacterium leprae</i> |

Phylogenetic relationships in each group are based on published phylogenies.

Table 3 HO-to-CD and CD-to-HO inversion rates based on the inversions inferred by the phylogeny-based standard method.

| Species | Current no. of CD genes | Current no. of HO genes | HO-to-CD inversions | CD-to-HO inversions | HO-to-CD inversion rate ^a | CD-to-HO inversion rate ^b | P value ^c |
|-----------------------------------|-------------------------|-------------------------|---------------------|---------------------|--------------------------------------|--------------------------------------|------------------------|
| <i>Mycoplasma gallisepticum</i> | 271 | 48 | 11 | 21 | 0.290 | 0.075 | 3.5 × 10 ⁻⁵ |
| <i>Staphylococcus aureus</i> | 1175 | 310 | 16 | 26 | 0.053 | 0.022 | 3.4 × 10 ⁻³ |
| <i>Bacillus subtilis</i> | 1646 | 406 | 8 | 30 | 0.021 | 0.019 | 0.71 |
| <i>Campylobacter jejuni</i> | 553 | 309 | 68 | 89 | 0.236 | 0.155 | 3.6 × 10 ⁻³ |
| <i>Klebsiella pneumoniae</i> | 1515 | 1169 | 3 | 9 | 0.0026 | 0.0059 | 0.20 |
| <i>Mycobacterium tuberculosis</i> | 930 | 450 | 8 | 22 | 0.018 | 0.023 | 0.56 |

^aNumber of HO-to-CD inversions divided by the number of previously HO genes, which is the number of current HO genes plus the number of HO-to-CD inversions minus the number of CD-to-HO inversions.

^bNumber of CD-to-HO inversions divided by the number of previously CD genes, which is the number of current CD genes plus the number of CD-to-HO inversions minus the number of HO-to-CD inversions.

^cChi-squared test of equal rates of HO-to-CD and CD-to-HO inversions, performed using a 2 × 2 table of the number of previously HO genes that remain HO, number of HO-to-CD inversions, number of previously CD genes that remain CD, and the number of CD-to-HO inversions. The P values have not been corrected for multiple testing.

protein-related selection^{7,10,11}. By contrast, Merrikh and Merrikh computed the GC skew of a gene using its entire coding region, further increasing the likelihood of erroneous inferences of gene inversion.

We thus used the standard method to infer gene inversion in the six species analyzed by Merrikh and Merrikh. For each focal species A, a closely related species B from the same genus and an outgroup species C were chosen (Table 2), and orthologs among the three species were identified by reciprocal best hits from protein BLAST analysis (Supplementary Data 1). Gene orientation was determined from reference genome assemblies and inversions were inferred using the parsimony principle. We then counted the number of inversions in the lineage leading to the

focal species A from the common ancestor of A and B. The results showed that the rate of CD-to-HO inversion is significantly different from the rate of HO-to-CD inversion in three of the six species. In all three cases, the former rate is lower than the latter rate (Table 3), again supporting the mutation-selection balance hypothesis but refuting the adaptive hypothesis. Under the assumption that the number of HO genes observed from a species has reached equilibrium, the number of CD-to-HO inversions should equal the number of HO-to-CD inversions. However, we observed that the former is greater than the latter in each species (Table 3). This is likely because the evolutionary time concerned (since the divergence of species A from B) is relatively short such that a sizable fraction of deleterious CD-to-HO inversions has yet

to be purged. Such a time lag in the effect of purifying selection is well known¹². It is notable that we identified 114 HO-to-CD and 197 CD-to-HO inversions in the six species, while Merrikh and Merrikh identified an order of magnitude more inversions (1468 and 3669, respectively) in the same species². Apart from the general unreliability of Merrikh and Merrikh's method, the large disparity may also be due to the fact that we aimed to detect inversions that occurred since the divergence between species A and B, while inversions detected by Merrikh and Merrikh could have occurred at any time (though older inversions have lower detectabilities). The fact that only 1 HO-to-CD and 145 CD-to-HO inversions we detected are also detected by Merrikh and Merrikh suggests that their method not only has potential false-positive errors but also makes numerous false-negative errors. The particularly high false-negative rate in detecting HO-to-CD inversions by Merrikh and Merrikh's method is probably because G is selectively favored over C on the coding strand due to the fact that G-containing codons tend to code for amino acids with relatively low synthetic costs⁹, a confounding factor ignored in Merrikh and Merrikh's method. Another possibility is that, if a gene of species A has been inverted twice, once shortly before and once after the separation of A from B, the standard method will recognize the more recent inversion, but Merrikh and Merrikh's method will likely miss it.

In addition to analyzing gene inversions, Merrikh and Merrikh estimated the synonymous (d_S) and nonsynonymous (d_N) nucleotide substitution rates of individual genes². They reported that d_S is not significantly different between HO and CD genes, but d_N and d_N/d_S are significantly higher for HO than CD genes. The d_S comparison suggests that the point mutation rate is not different between HO and CD genes in coding regions, consistent with experimental data^{13,14} and arguing squarely against the basis of the adaptive hypothesis that the (beneficial) mutation rate is higher for HO than CD genes. Merrikh and Merrikh interpreted the results on d_N and d_N/d_S as evidence for positive selection of HO genes. However, higher d_N and d_N/d_S could also result from a relaxation of purifying selection. Given that highly expressed genes and essential genes are underrepresented among HO genes, relaxation of purifying selection seems a more reasonable interpretation¹⁵. Similarly, the observation of a larger fraction of genes with $d_N/d_S > 1$ among HO genes than CD genes² can be explained by a relaxation of purifying selection on HO genes, because no statistical test was performed by Merrikh and Merrikh to show that any gene has its d_N/d_S significantly exceeding 1, the criterion for establishing positive selection. If such statistical tests are to be performed, corrections for multiple testing would be necessary to guard against false-positive results. Merrikh and Merrikh² also reported enrichment of several functional groups among HO genes relative to CD genes. This non-randomness could be a byproduct of the known differences between HO and CD genes in expression level and essentiality^{5,6}, so cannot be used to support the adaptive hypothesis.

In conclusion, our reanalysis of the empirical data of Merrikh and Merrikh² and new analysis found no evidence for the adaptive hypothesis of lagging-strand encoding in bacterial genomes. Instead, all available data are broadly consistent with the mutation-selection balance hypothesis.

Methods

Data sources. All genome and sequence data of the species used in this study (see Table 2) were downloaded from NCBI (<ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/>). The reanalysis presented in Table 1 was based on the data in Supplementary Table 1 of Merrikh and Merrikh².

Estimation of inversion rates. The HO-to-CD inversion rate was estimated by the number of HO-to-CD inversion events divided by the number of previously HO genes. The number of previously HO genes is the number of present-day HO genes

plus the number of HO-to-CD inversion events minus the number of CD-to-HO inversion events. The CD-to-HO inversion rate was similarly estimated.

Phylogeny-based identification of gene inversions. For each focal species A, a closely related species (B) from the same genus and an outgroup species (C) were chosen. In each three-species group, protein BLAST analysis was performed to identify orthologs. Gene orientations were determined from reference genome assemblies, followed by the inference of inversion events using the parsimony principle. Specifically, in a three-species orthologous gene group, when the gene orientation is different between the two ingroup species, an inversion event is inferred for the ingroup species in which the gene orientation differs from that in the outgroup.

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

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files, as well as from the NCBI Reference Sequence Database (<ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/>). Accessions are listed in Supplementary Data 1.

Received: 14 January 2020; Accepted: 12 April 2021;
Published online: 12 May 2022

References

- Merrikh, H., Zhang, Y., Grossman, A. D. & Wang, J. D. Replication-transcription conflicts in bacteria. *Nat. Rev. Microbiol.* **10**, 449–458 (2012).
- Merrikh, C. N. & Merrikh, H. Gene inversion potentiates bacterial evolvability and virulence. *Nat. Commun.* **9**, 4662 (2018).
- Chen, X. & Zhang, J. Why are genes encoded on the lagging strand of the bacterial genome? *Genome Biol. Evol.* **5**, 2436–2439 (2013).
- Lynch, M. et al. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).
- Brewer, B. J. When polymerases collide: replication and the transcriptional organization of the E. coli chromosome. *Cell* **53**, 679–686 (1988).
- Rocha, E. P. & Danchin, A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* **31**, 6570–6577 (2003).
- Mackiewicz, P. et al. The differential killing of genes by inversions in prokaryotic genomes. *J. Mol. Evol.* **53**, 615–621 (2001).
- Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660–665 (1996).
- Chen, W. H., Lu, G., Bork, P., Hu, S. & Lercher, M. J. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat. Commun.* **7**, 11334 (2016).
- Tillier, E. R. & Collins, R. A. Replication orientation affects the rate and direction of bacterial gene evolution. *J. Mol. Evol.* **51**, 459–463 (2000).
- Fonseca, M. M., Posada, D. & Harris, D. J. Inverted replication of vertebrate mitochondria. *Mol. Biol. Evol.* **25**, 805–808 (2008).
- Rocha, E. P. et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* **239**, 226–235 (2006).
- Schroeder, J. W., Hirst, W. G., Szweczyk, G. A. & Simmons, L. A. The effect of local sequence context on mutational bias of genes encoded on the leading and lagging strands. *Curr. Biol.* **26**, 692–697 (2016).
- Sankar, T. S., Wastuwidyanyingtyas, B. D., Dong, Y., Lewis, S. A. & Wang, J. D. The nature of mutations induced by replication–transcription collisions. *Nature* **535**, 178–181 (2016).
- Zhang, J. & Yang, J. R. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **16**, 409–420 (2015).

Acknowledgements

This work was supported in part by the US National Institutes of Health research grant R35GM139484 to J.Z.

Author contributions

H.L. and J.Z. designed the research and wrote the paper. H.L. conducted the research and analyzed the data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-30000-8>.

Correspondence and requests for materials should be addressed to Jianzhi Zhang.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022