AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

OBSERVATION
Ecological and Evolutionary Science

An Important Role for Purifying Selection in Archaeal Genome Evolution

Zhe Lyu,[a,b] Zhi-Gang Li,[c,d] Fei He,[c] Ziding Zhang[c]

College of Resources and Environmental Sciences, China Agricultural University, Beijing, China[a]; Department of Microbiology, University of Georgia, Athens, Georgia, USA[b]; State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, China[c]; State Key Laboratory of Agrobiotechnology and Ministry of Agriculture Key Laboratory of Plant Pathology, Beijing, China[d]

ABSTRACT As the null hypothesis of genome evolution, population genetic theory suggests that selection strength controls genome size. Through the process of genetic drift, this theory predicts that compact genomes are maintained by strong purifying selection while complex genomes are enabled by weak purifying selection. It offers a unifying framework that explains why prokaryotic genomes are much smaller than their eukaryotic counterparts. However, recent findings suggest that bigger prokaryotic genomes appear to experience stronger purifying selection, indicating that purifying selection may not dominate prokaryotic genome evolution. Since archaeal genomes were underrepresented in those studies, generalization of the conclusions to both archaeal and bacterial genomes may not be warranted. In this study, we revisited this matter by focusing on archaeal and bacterial genomes separately. We found that bigger bacterial genomes indeed experienced stronger purifying selection, but the opposite was observed in archaeal genomes. This new finding would predict an enrichment of noncoding sequences in large archaeal genomes, which was confirmed by an analysis of coding density. In contrast, coding density remained stable regardless of bacterial genome size. In conclusion, this study suggests that purifying selection may play a more important role in archaeal genome evolution than previously hypothesized, indicating that there could be a major difference between the evolutionary regimes of *Archaea* and *Bacteria*.

IMPORTANCE The evolution of genome complexity is a fundamental question in biology. A hallmark of eukaryotic genome complexity is that larger genomes tend to have more noncoding sequences, which are believed to be minimal in archaeal and bacterial genomes. However, we found that archaeal genomes also possessed this eukaryotic feature while bacterial genomes did not. This could be predicted from our analysis on genetic drift, which showed a relaxation of purifying selection in larger archaeal genomes, also a eukaryotic feature. In contrast, the opposite was evident in bacterial genomes.

KEYWORDS *Archaea*, evolution, genome analysis

Received 22 August 2017 **Accepted** 6 October 2017 **Published** 24 October 2017

Citation Lyu Z, Li Z-G, He F, Zhang Z. 2017. An important role for purifying selection in archaeal genome evolution. mSystems 2: e00112-17. https://doi.org/10.1128/mSystems.00112-17.

Editor Michael Rust, Institute for Genomics & Systems Biology

Address correspondence to Zhe Lyu, zhelyu@uga.edu, or Ziding Zhang, zidingzhang@cau.edu.cn.

Z.L. and Z.-G.L. contributed equally to this work.

Eukaryotic genomes vary in size by orders of magnitude more than prokaryotic genomes. The genome size range is about 10^6 to 10^{11} bp in eukaryotes, which are rich in noncoding sequences, and 10^5 to 10^7 bp in prokaryotes, which are streamlined and have minimal noncoding sequences (1, 2). This genome size gap is believed to be shaped primarily by nonadaptive processes based on the population genetic theory (3). That theory suggests that prokaryotes often undergo strong purifying selection owing to a generally large effective population size to maintain compact genomes (2). In contrast, eukaryotes typically have a much smaller effective population size and are subject to weak purifying selection, which enables large genomes (4). This is because all excess DNA is mutationally hazardous, and the efficiency of selection determines

whether the excess DNA is removed from or fixed in the genome through the process of genetic drift (5).

The efficiency of selection can be approximately measured by the genome-wide *dN*/*dS* ratio (ratio of nonsynonymous to synonymous substitution rates) for orthologous genes shared by closely related lineages, and the stronger the purifying selection becomes, the lower the *dN*/*dS* ratio is (6). The population genetic theory predicts that bigger genomes experience weaker purifying selection or higher *dN*/*dS* ratios (2–4). However, recent findings regarding prokaryotic genomes suggest otherwise, showing that genome size is negatively correlated with the *dN*/*dS* ratio (7, 8). This only seems possible when gene gains could be slightly beneficial and that the benefits would be diluted out because of genome expansion and deletion bias (i.e., DNA loss outpaces gain), based on mathematical models (8). The benefits of gene gains thus make genome expansion possible under strong purifying selection, but the expansion stops once the benefits diminish, thus restraining the overall genome size. Indeed, deletion bias appears to be universal across the full range of cellular life forms, and its strength tends to decline when the genome expands, indicating a dynamic balance between DNA loss and DNA gain (1, 9, 10).

While bacterial species were extensively sampled in these studies, archaeal species were underrepresented (7, 8). Therefore, generalization of the mechanisms identified therein may not be warranted in *Archaea*. In this study, we concentrated on archaeal genomes to revisit previous hypotheses of genome size evolution. Bacterial genomes were also sampled and examined for comparison when necessary. We observed that the strength of purifying selection and the amount of noncoding genes were negatively and positively associated with archaeal genome size, respectively, as predicted by population genetic theory. In contrast, the opposite trend was evident in bacterial genomes.

Expansion of archaeal genomes associated with relaxed purifying selection. In eukaryotes, relaxed purifying selection is associated with genome expansion, which is consistent with the accumulation of introns and mobile elements that are often deleterious (3).

While our bacterial data set reproduced the previous observations showing that strong purifying selection was associated with genome expansion (7, 8), our archaeal data set revealed a eukaryote-like pattern based on a genome-wide *dN*/*dS* ratio analysis (Fig. 1A and B; see Tables S1 and S2 in the supplemental material). This finding predicts an enrichment of noncoding sequences in larger archaeal genomes, as also observed in eukaryotic genomes (see below for coding density analysis).

At least a couple of observations indicated that our archaeal and bacterial data sets were comparable and likely representative. First, both archaeal and bacterial mean *dN*/*dS* ratios were between 0.05 and 0.20 (Fig. 1A and B) and a similar range has also been observed in eukaryotic lineages (11, 12). Second, most archaeal and bacterial genome pairs had an average nucleotide identity (ANI) of 75 to 95% (Fig. 2), an empirical range that delineated closely related prokaryotic species belonging to the same genus (13). This minimized the potential bias of *dN*/*dS* ratio computations caused by uneven variation in evolutionary distances between genomes (7). Nevertheless, while the bacterial samplings both here and in previous studies seem to have captured substantial taxonomic diversities of bacterial genomes, the archaeal samplings need further improvements. Sampling of archaeal genomes across more phyla will be necessary to begin to address their full genome size range. This is because the archaeal genome size narrowly ranges from 0.5 to 6 Mb compared to the 0.6- to 9-Mb range of bacterial genomes (Tables S3 and S4). Of the 29 prokaryotic phyla taxonomically established, only 2 belong to *Archaea*, where most complete archaeal genomes currently come from, and an estimated ~300 archaeal phyla still wait to be sampled (14).

Coding density in *Archaea* shows a trend similar to that seen in eukaryotes. Coding density or gene density, as measured by the proportion of a genome sequence that is composed of coding or gene sequence, has been shown to correlate negatively and neutrally with eukaryotic and bacterial genome sizes, respectively (7, 15, 16). This

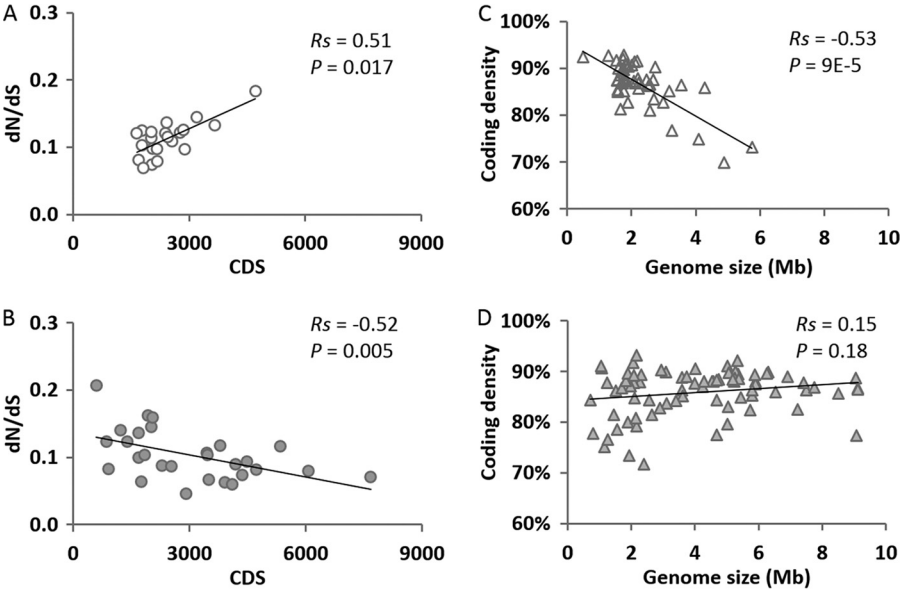

FIG 1 Association between genome size and the *dN*/*dS* ratio for archaeal (A; $n = 21$) and bacterial (B; $n = 28$) genome pairs and association between coding density and genome size in *Archaea* (C; $n = 49$) and *Bacteria* (D; $n = 78$). Note that the stronger the purifying selection becomes, the lower the *dN*/*dS* ratio is. Spearman rank correlation coefficients (*Rs*) and two-sided significance (*P*) values are indicated in each panel. Genome sizes are shown in megabase pairs (Mb) in panels C and D but are in protein coding sequences (CDS) in panels A and B, as only protein coding sequences were used for *dN*/*dS* ratio analysis. Regardless, the same trends were reproduced when genome size measured in base pairs instead of protein coding sequences was used.

observation made in eukaryotes is consistent with a genomic structure showing that noncoding sequences, primarily spliceosomal introns and mobile genetic elements, are overrepresented in larger eukaryotic genomes (3). The observation regarding *Bacteria* is also consistent with the dynamic balance between DNA loss and DNA gain (1, 8, 9). While our bacterial data set confirmed previous observations, our archaeal data set again revealed a eukaryote-like pattern (Fig. 1C and D and Tables S3 and S4). That is to say, a strong negative correlation was observed between coding density and archaeal genome size, indicating that an insertion bias enriching noncoding sequences was also evident during the expansion of archaeal genomes. It remains elusive what category of noncoding sequences is overrepresented in larger archaeal genomes, as the nature of most archaeal noncoding sequences is poorly characterized (17). Regardless of the category, it would probably be different from that observed in eukaryotes, as spliceosomal introns have never been found in *Archaea* and the distribution of known mobile elements in archaeal genomes is similar to that in their bacterial counterparts (18). However, the presence of novel mobile elements in *Archaea* could not be ruled out. Compared to eukaryotes, one common feature is still shared by *Archaea* and *Bacteria*, both of which had a similar and narrow range of coding density, i.e., about 70 to 95%

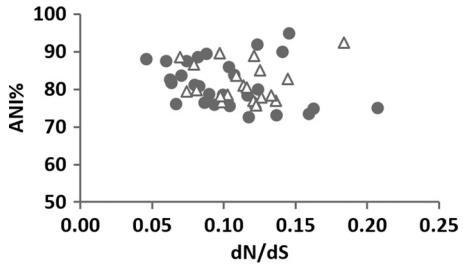

FIG 2 Association between the *dN*/*dS* ratio and genomic distance as measured by percent ANI. Closed circles represent bacterial pairs ($n = 28$), and open triangles represent archaeal pairs ($n = 21$).

in our data set. In contrast, the eukaryotic coding density roughly ranged from 1 to 80% (15). Therefore, there could be an unknown force(s) that prevents noncoding regions in archaeal genomes from enriching to the extremely high levels observed in eukaryotes.

This study presents the first evidence that the evolutionary regimen of the complexity of archaeal genomes could be significantly different from that of their bacterial counterparts. Instead, certain key eukaryote-like evolutionary features seem to be already embedded in archaeal genomes. On the one hand, those observations seem striking, as the architecture of archaeal genomes is very similar to that of bacterial genomes (2). On the other hand, they may simply reflect the close evolutionary connections between *Archaea* and eukaryotes, since eukaryotes likely evolved within the *Archaea* (19). Although the archaeal samplings here are still limited and the similarities drawn between *Archaea* and eukaryotes here are preliminary, we believe our study serves as a sufficient reminder that it is probably no longer a safe practice to group archaeal and bacterial genomes into one category when testing evolutionary hypotheses. Rather, comparisons should be made among *Archaea*, *Bacteria*, and eukaryotes.

Data collection and coding density calculation. Genomic data from closely related species that belong to the same genus were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/). Genera or species that were not well established, i.e., those whose taxonomic names were not formally proposed or alternative names were available, were removed before further analysis. The coding density of each genome was calculated as follows: % coding density = (DNA length of all coding sequences/total DNA length of complete genome) $\times$100%.

***dN/dS* ratios and ANI analyses.** For each pair of genomes, pairs of orthologs were identified by a hybrid procedure combining Bidirectional Best Hit and BLASTclust (20). The orthologous pairs identified were aligned by ClustalW2, and *dN*/*dS* ratios were calculated by using YN00 in the PAML package (21, 22). Synonymous sites and ratios of synonymous to nonsynonymous sites considered to be saturated or unreliable (that is, a *dS* of $<$0.1, a *dS* of $>$1.6, or a *dN*/*dS* ratio of $>$99) were discarded before the mean *dN*/*dS* ratio and standard error of the mean of each genome pair were both calculated and reported in Tables S1 and S2. Although data filtration is a common practice to ensure better estimation of selective pressure, it also substantially reduces the number of pairwise orthologs available for downstream analyses (7, 11). This would make the data set less representative for inferring genome-wide patterns. Therefore, only genome pairs retaining pairwise orthologs accounting for no less than 5% of the pairs' average coding capacity after all filtrations were kept for ultimate analyses. This cutoff is the starting point to represent most of the essential genes determined in prokaryotic genomes (23). ANI values were retrieved from IMG (24).

Correlation analysis. Spearman rank correlation (v1.0.3) was employed to obtain correlation coefficients (*Rs*) and two-sided *P* values (25).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00112-17.

TABLE S1, DOCX file, 0.05 MB.

TABLE S2, DOCX file, 0.03 MB.

TABLE S3, DOCX file, 0.04 MB.

TABLE S4, DOCX file, 0.04 MB.

ACKNOWLEDGMENTS

We thank William B. Whitman, Deng-Ke Niu, and anonymous reviewers for suggestions about this work. We are in debt to Yahai Lu for strong support with the initiation of the long-term cooperation among the coauthors. We apologize for being unable to cite all of the relevant references because of space limitation.

This work was supported in part by National Natural Science Foundation of China grant 31471249.

REFERENCES

1. Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596. https://doi.org/10.1016/S0168-9525(01)02447-7.
2. Lynch M. 2006. Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol 60:327–349. https://doi.org/10.1146/annurev.micro.60.080805.142300.
3. Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404. https://doi.org/10.1126/science.1089370.
4. Lynch M. 2006. The origins of eukaryotic gene structure. Mol Biol Evol 23:450–468. https://doi.org/10.1093/molbev/msj050.
5. Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A 112:15690–15695. https://doi.org/10.1073/pnas.1514974112.
6. Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503. https://doi.org/10.1016/S0169-5347(00)01994-7.
7. Novichkov PS, Wolf YI, Dubchak I, Koonin EV. 2009. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol 191:65–73. https://doi.org/10.1128/JB.01237-08.
8. Sela I, Wolf YI, Koonin EV. 2016. Theory of prokaryotic genome evolution. Proc Natl Acad Sci U S A 113:11399–11407. https://doi.org/10.1073/pnas.1614083113.
9. Kuo CH, Ochman H. 2009. Deletional bias across the three domains of life. Genome Biol Evol 1:145–152. https://doi.org/10.1093/gbe/evp016.
10. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. Science 287:1060–1062. https://doi.org/10.1126/science.287.5455.1060.
11. Kelkar YD, Ochman H. 2012. Causes and consequences of genome expansion in fungi. Genome Biol Evol 4:13–23. https://doi.org/10.1093/gbe/evr124.
12. Hutter B, Bieg M, Helms V, Paulsen M. 2010. Divergence of imprinted genes during mammalian evolution. BMC Evol Biol 10:116. https://doi.org/10.1186/1471-2148-10-116.
13. Kim M, Oh HS, Park SC, Chun J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol 64:346–351. https://doi.org/10.1099/ijs.0.059774-0.
14. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12:635–645. https://doi.org/10.1038/nrmicro3330.
15. Hou Y, Lin S. 2009. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. PLoS One 4:e6978. https://doi.org/10.1371/journal.pone.0006978.
16. Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res 19:1450–1454. https://doi.org/10.1101/gr.091785.109.
17. Makarova KS, Wolf YI, Forterre P, Prangishvili D, Krupovic M, Koonin EV. 2014. Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. Extremophiles 18:877–893. https://doi.org/10.1007/s00792-014-0672-7.
18. Filée J, Siguier P, Chandler M. 2007. Insertion sequence diversity in archaea. Microbiol Mol Biol Rev 71:121–157. https://doi.org/10.1128/MMBR.00031-06.
19. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T, Banfield JF, Schramm A, Baker BJ, Spang A, Ettema TJG. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358. https://doi.org/10.1038/nature21031.
20. Madan Babu M, Teichmann SA, Aravind L. 2006. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 358:614–633. https://doi.org/10.1016/j.jmb.2006.02.019.
21. Friedman R, Drake JW, Hughes AL. 2004. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. Genetics 167:1507–1512. https://doi.org/10.1534/genetics.104.026344.
22. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088.
23. Luo H, Lin Y, Gao F, Zhang CT, Zhang R. 2014. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res 42:D574–D580. https://doi.org/10.1093/nar/gkt1131.
24. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A. 2015. Microbial species delineation using whole genome sequences. Nucleic Acids Res 43:6761–6771. https://doi.org/10.1093/nar/gkv657.
25. Wessa P. 2017. Spearman rank correlation (v1.0.3) in free statistics software (v1.2.1). Office for Research Development and Education, Leuven Institute for Research on Information Systems, University of Leuven, Leuven, Belgium. https://www.wessa.net/rwasp_spearman.wasp/.