nature ecology & evolution

BRIEF COMMUNICATION

https://doi.org/10.1038/s41559-017-0425-y

Evolutionary determinants of genome-wide nucleotide composition

Hongan Long[1], Way Sung[2], Sibel Kucukyildirim[3], Emily Williams[4], Samuel F. Miller[5], Wanfeng Guo[4], Caitlyn Patterson[5], Colin Gregory[5], Chloe Strauss[5], Casey Stone[5], Cécile Berne[5], David Kysela[5], William R. Shoemaker[5], Mario E. Muscarella[6], Haiwei Luo[7], Jay T. Lennon[5], Yves V. Brun[5] and Michael Lynch[4*]

One of the long-standing mysteries of evolutionary genomics is the source of the wide phylogenetic diversity in genome nucleotide composition (G + C versus A + T), which must be a consequence of interspecific differences in mutation bias, the efficiency of selection for different nucleotides or a combination of the two. We demonstrate that although genomic G + C composition is strongly driven by mutation bias, it is also substantially modified by direct selection and/or as a by-product of biased gene conversion. Moreover, G + C composition at fourfold redundant sites is consistently elevated above the neutral expectation—more so than for any other class of sites.

For some classes of genomic sites, G + C nucleotide composition covers nearly the full range of possible variation (frequencies of ~0.0 to ~1.0) across species[1–5]. It is commonly thought that the contribution of mutation to such variation can be determined from the nucleotide content of fourfold redundant (synonymous) sites within codons or from the composition of rare variants, and analyses of this type have led to the idea that mutation is universally biased in the direction of A + T[6–8]. However, selection on such sites can bias such interpretations. To eliminate such issues, we use direct estimates of the mutation spectra derived from mutation-accumulation experiments and/or parent–offspring trios for 37 diverse species.

Of the datasets analysed herein, 25 involve published data (summarized in ref. [9] with respect to mutation rates) and 12 involve long-term mutation-accumulation experiments in diverse microbial species reported here for the first time (Supplementary Tables 1–3). Each new mutation-accumulation experiment involves the complete genome sequencing of ~50 lines serially transferred through single-cell bottlenecks for thousands of cell divisions, which effectively eliminates the ability of natural selection to significantly modify the accumulation of all but the small fraction of extremely deleterious mutations (which in any case are irrelevant to the following analyses, as they do not accumulate evolutionarily; ref. [9]). From the resultant spectra for base-substitution mutations (typically based on dozens to hundreds of de novo mutations), letting m be the ratio of the per-nucleotide mutation rate in the G + C → A + T direction to the reciprocal rate, the expected equilibrium G + C composition under neutrality (where mutation is the only directional evolutionary force) is

$$\tilde{P}_n = \frac{1}{1+m} \quad (1)$$

Comparison of the observed genome-wide nucleotide compositions of the study species with these neutral expectations reveals several general patterns (Fig. 1). First, mutation biases in unicellular species may be in either the A + T or G + C direction (leading to a $\tilde{P}_n$ value less than or greater than 0.5, respectively), although the former is most common, and no characterized multicellular eukaryote has mutation bias in the G + C direction. Second, regardless of the class of DNA or the phylogenetic grouping, with few exceptions, genome-wide G + C composition is close to or substantially above the neutral expectation, implying the existence of a near-universal directional force favouring G + C content. Third, the primary exception to this pattern involves zerofold redundant sites (where all nucleotide substitutions lead to amino-acid changes) in bacteria with endogenous mutation pressure towards G + C ($\tilde{P}_n > 0.5$), where selection for amino acids containing A + T in such codon positions apparently takes precedence over other G + C enhancing forces. This tendency is reflected in the diminished slope in the regression involving such sites (Supplementary Table 4). Fourth, for two- and fourfold redundant sites (where two and four nucleotides respectively encode for the same amino acid), G + C composition is particularly strongly elevated, by an average amount that is essentially independent of the neutral expectation, but with considerable variation. The strong elevation for fourfold redundant sites implies the existence of general forces favouring G + C independent of the implications for the proteome.

The magnitude of the strength of selection required to account for the deviation of G + C composition at fourfold redundant sites relative to the neutral expectation can be estimated by noting that in the presence of selection, equation (1) generalizes to

$$P_s = \frac{1}{1+me^{-S}} \quad (2)$$

[1]Institute of Evolution and Marine Biodiversity, KLMME, Ocean University of China, Qingdao, China. [2]Department of Bioinformatics and Genomics, University of North Carolina, Charlotte, NC, USA. [3]Department of Biology, Hacettepe University, Ankara, Turkey. [4]Center for Mechanisms of Evolution, Arizona State University, PO Box 877701, Tempe, AZ, USA. [5]Department of Biology, Indiana University, Bloomington, IN, USA. [6]Department of Plant Biology, University of Illinois, Urbana–Champaign, Champaign, IL, USA. [7]School of Life Sciences and Partner State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong. Hongan Long, Way Sung and Sibel Kucukyildirim contributed equally to this work.
*e-mail: mlynch11@asu.edu

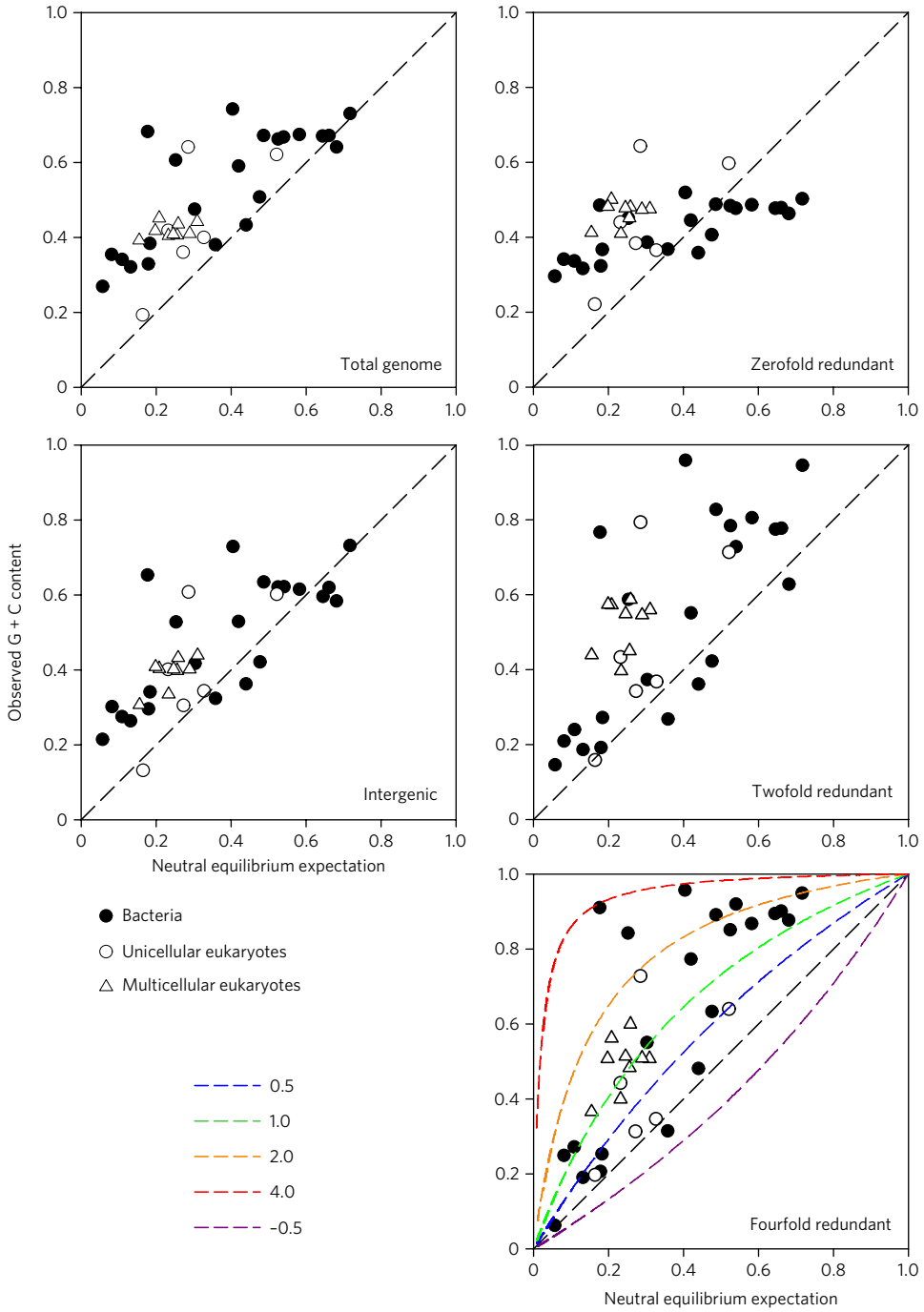

Fig. 1 | Relationship between genome-wide nucleotide composition and the neutral expectation. The data are subdivided into three major groups of organisms. The diagonal dashed lines denote agreement with the neutral expectation, with points above the diagonal reflecting conditions in which there is selection for elevated G + C content. For reference, the lower panel provides isoclines of expected genome compositions under selection, with values of the composite parameter $S = \phi N_e s$ being equivalent to the ratio of the power of selection in favour of G + C content relative to the power of genetic drift. The neutral equilibrium expectation is calculated from equation (1) and the observed G + C content is based on direct observation of genome contents. All data can be found in Supplementary Tables 1–5.

where $S = \phi N_e s$, with N_e being the effective population size, $\phi = 2$ or 4 for haploids and diploids respectively, and s being the average selective advantage of G + C nucleotides over A + T[10,11]. S for each genomic category is shown in Supplementary Table 5. In Fig. 1, lines of expectation for P_s for various values of S (equivalent to the ratio of the power of selection s to the power of drift $1/(\phi N_e)$) show that S (in favour of G + C) is generally in the range 0.5–4.0. Thus, some selective force in favour of G + C composition is pervasive and relatively strong, although not strong enough to entirely overcome the mutational expectations.

The results for fourfold redundant sites are of relevance to the common usage of measures of standing variation at such positions to estimate N_e under the assumption of neutrality (drift-mutation equilibrium), which leads to an expected average heterozygosity of $\tilde{\pi} \approx \phi N_e u$, where u is the mean mutation rate per nucleotide site. From a rearrangement of equation (15) in ref. [12], the ratio of

heterozygosity under drift-mutation-selection equilibrium and the neutral expectation is

$$\frac{\pi_s}{\tilde{\pi}_n} = \frac{(1+m)(e^S-1)}{S(m+e^S)} \quad (3)$$

the solution of which shows that when mutation is strongly biased towards A+T but selection strongly favours G+C, the expected nucleotide diversity can be several-fold greater than the neutral expectation ($\tilde{\pi}$), which would lead to the same proportional overestimation of N_e when the mutation rate is factored out (Fig. 2). When mutation bias and selection operate in the same direction, π can be downwardly biased up to a few-fold with respect to the neutral condition. Thus, relative estimates of N_e derived from silent-site variation can be off by several fold (when compared with each other) if selection is moderately strong and there are strong differences in mutation bias among contrasted species, which, based on the wide range of estimated $\tilde{P}_n$, is clearly the case.

Our results imply a near-universal pervasive mechanism operating to increase G+C content, as previously inferred indirectly from polymorphism data for G+C-rich genomes[7]. However, the sources of such selection remain unclear. Given that the substantial number of species in this study inhabit a wide range of environments and are derived from a diversity of bacterial and eukaryotic lineages, consistent directional selection in favour of G+C is not readily reconciled by ecological and/or genetic-background arguments. Moreover, given that such selection is experienced by both silent and replacement codon sites, arguments based on protein-sequence constraints and transcription fidelity are not compelling. Likewise, because the pattern extends to intergenic (largely non-coding) DNA, arguments based on gene expression and translation speed and accuracy[13,14] do not seem to apply. Although gene expression levels within species are correlated with local gene G+C composition, all but one r^2 values involving these variables are ≤ 0.02, and the signs of the relationships are inconsistent (Supplementary Table 6). One general force that may be of relevance is DNA stability, in that G:C pairs involve three hydrogen bonds, whereas A:T pairs involve only two.

An alternative explanation for near-universal pressure towards G+C content involves gene conversion, which results from the repair of heteroduplex DNA arising from recombination between two non-identical sequences and, if biased, can operate like selection at the population-genetic level. In every eukaryote and bacterium that has been closely scrutinized, gene conversion has been found to be biased in the direction of G+C (refs [15–21]), although the molecular mechanisms encouraging such universal behaviour are unknown. Most attempts to estimate S associated with codon bias (which may be driven by biased gene conversion) have yielded estimates on the order of 0.1 to 4.0 in diverse phylogenetic groups[4] (although not always in the G+C direction) and our results (Fig. 1) are fully compatible with this magnitude of selection.

Because effective population sizes vary among organisms by several orders of magnitude, this small range in S suggests that there must be a roughly inverse relationship between N_e and s, whatever the force encouraging G+C content. Under a scenario of natural selection, such a condition is expected under any concave fitness function for increasing G+C content, as the selective advantage of incremental changes would then diminish with increased G+C (further out on the fitness plateau) and larger population sizes would enhance the efficiency of selection for higher G+C content. However, a scenario of biased gene conversion requires a rather different set of conditions—the magnitude of the biasing force (towards G+C) would have to increase with decreasing N_e. In principle, this might occur if a large fraction of G+C conversions were deleterious, as natural selection opposing conversion-driven G+C would be reduced in the face of increased random genetic drift[9]. This would, however, also require a very strong increase in the biasing force in small populations because biased gene conversion depends on both the asymmetric force and the recombination rate per nucleotide site, with the recombination rate actually scaling negatively with N_e (ref. [4]).

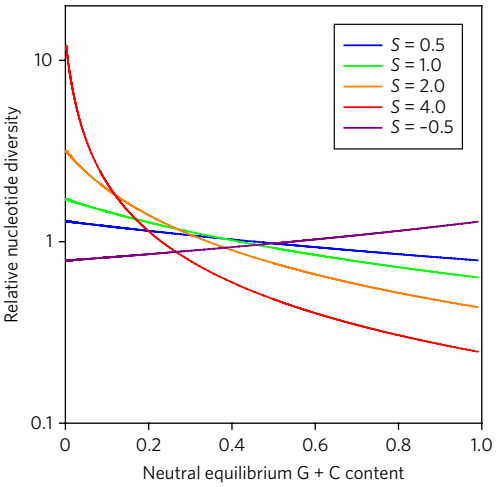

Fig. 2 | Expected equilibrium levels of within-population nucleotide diversity scaled by the neutral expectation. Derived from equation (3) in the text, with various strengths of selection (S) colour coded as in Fig. 1.

In summary, our results conclusively support the idea that genome-wide nucleotide composition is strongly influenced by mutation bias at all classes of sites, but that phylogenetically general directional forces beyond mutation (natural selection and/or biased gene conversion) play a role as well. The positive association between neutral G+C composition expectations and the actual utilization at zerofold redundant sites demonstrates that even amino-acid usage is dictated at least in part by mutation pressure, with the G+C content of such sites differing more than twofold between genomes with strong mutation bias towards A+T versus those with bias towards G+C (Fig. 1 and Supplementary Table 1). However, despite this gradient, G+C utilization at zerofold redundant sites is generally substantially greater than the neutral expectation when the latter is <0.5, so the possibility that such content is influenced by the same selection pressures favouring G+C content at silent sites cannot be ruled out.

Finally, although the ultimate sources of variation in the mutation spectrum (which drives the wide range of variation in nucleotide composition among species) are unknown, they may involve effectively neutral processes. Owing to the predominance of deleterious mutations, selection is expected to generally drive the genome-wide mutation rate down to some level beyond which further advantages are offset by the power of random genetic drift[9]. However, any particular mutation rate can be compatible with a wide range of mutational spectra, which may be free to wander over evolutionary time, conditional on the maintenance of a constant genome-wide deleterious rate[22]. Notably, when the prevailing mutation pressure towards A+T is in conflict with the forces favouring G+C content (which is true for most taxa), the average genome-wide mutation rate per nucleotide site is indirectly inflated, owing to the elevated abundance of more mutable (G and C) nucleotides.

Methods

G+C composition calculation. Mutation spectra, strain culturing and reference-genome information for the 37 species in this study are presented in Supplementary Table 1. We enumerated all sites of the genomes to calculate genome-wide G+C nucleotide compositions. For the G+C nucleotide composition at different functional sites of the genomes, we parsed out: (1) the second nucleotides of all codons except stop codons to delineate zerofold

redundant sites; (2) the third nucleotides of codons for Asn, Asp, Cys, Gln, Glu, His, Lys and Tyr amino acids for twofold redundant sites; (3) the third nucleotides of codons for Ala, Arg, Gly, Leu, Pro, Ser, Thr and Val amino acids for fourfold redundant sites; and (4) the nucleotides between the start and/or stop codons (or between untranslated regions when annotated) of two adjacent genes for intergenic DNA. Expression data of each gene were downloaded from the National Center for Biotechnology Information Gene Expression Omnibus database and the gene-specific G+C contents for fourfold redundant sites were parsed as above. The statistical details of the relationship between gene expression and fourfold redundant site G+C composition are in Supplementary Table 6.

Mutation-accumulation line transfers. For the 12 new microbial mutation-accumulation projects reported in this study, all lines were cultured under ideal conditions on solid agar plates, using procedures relied on in numerous previous studies summarized in ref. [9]. Within each study, all mutation-accumulation lines initiated from a single-cell progenitor and were then single-cell transferred daily to weekly (depending on the growth rate; necessary for visual localization of the colonies). Each month, numbers of cell divisions during each culturing cycle were estimated using colony-forming units from serial-dilution procedures.

Genome sequencing and raw data. Genomic DNA of the mutation-accumulation lines was extracted using the Wizard Genomic DNA Purification Kit (Promega). Illumina libraries for genome sequencing were then constructed using an optimized Nextera DNA Library Prep Kit (Illumina), and 150- or 250-base pair paired-end Illumina sequencing was performed on a HiSeq 2500 platform (Hubbard Center for Genome Studies, University of New Hampshire). Read trimming, mapping and mutation rate calculations followed ref. [23]. Duplicate reads were removed using picard-tools-2.5.0 (https://github.com/broadinstitute/picard) in GATK 3.6. Unique single nucleotide polymorphism and indel variants were analysed with HaplotypeCaller and standard hard-filtering parameters described by GATK Best Practices recommendations[24–26]. Candidate variants were identified visually with the Integrated Genome Viewer (version 2.3.5)[27]. All base substitutions, insertions and deletions identified are in Supplementary Tables 2 and 3.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. Raw reads of genome sequencing generated in this study are available in the National Center for Biotechnology Information Sequence Read Archive with BioProject number PRJNA376572.

Received: 21 May 2017; Accepted: 21 November 2017;
Published online: 1 January 2018

References

1. Sueoka, N. *Proc. Natl Acad. Sci. USA* **48**, 582–592 (1962).
2. Gu, X., Hewett-Emmett, D. & Li, W. H. *Genetica* **102–103**, 383–391 (1998).
3. Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L. & McAdams, H. H. *Proc. Natl Acad. Sci. USA* **101**, 3480–3485 (2004).
4. Lynch, M. The Origin of Genome Architecture. Sinauer Associates: Sunderland, 2007.
5. Rocha, E. P. C. & Feil, E. J. *PLoS Genet.* **6**, e1001104 (2010).
6. Hershberg, R. & Petrov, D. A. *PLoS Genet.* **6**, e1001115 (2010).
7. Hildebrand, F., Meyer, A. & Eyre-Walker, A. *PLoS Genet.* **6**, e1001107 (2010).
8. Lynch, M. *Proc. Natl Acad. Sci. USA* **107**, 961–968 (2010).
9. Lynch, M. et al. *Nat. Rev. Genet.* **17**, 704–714 (2016).
10. Li, W. H. *J. Mol. Evol.* **24**, 337–345 (1987).
11. Bulmer, M. *Genetics* **129**, 897–907 (1991).
12. McVean, G. A. T. & Charlesworth, B. *Genet. Res.* **74**, 145–158 (1999).
13. Raghavan, R., Kelkar, Y. D. & Ochman, H. *Proc. Natl Acad. Sci. USA* **109**, 14504–14507 (2012).
14. Kelkar, Y. D., Phillips, D. S. & Ochman, H. *G3* **5**, 1247–1252 (2015).
15. Marais, G., Mouchiroud, D. & Duret, L. *Genet. Res.* **81**, 79–87 (2003).
16. Mancera, E., Bourgon, R., Brozzi, A., Huber, W. & Steinmetz, L. M. *Nature* **454**, 479–485 (2008).
17. Galtier, N., Duret, L., Glemin, S. & Ranwez, V. *Trends Genet.* **25**, 1–5 (2009).
18. Pessia, E. et al. *Genome Biol. Evol.* **4**, 675–682 (2012).
19. Williams, A. L. et al. *eLife* **4**, e04637 (2015).
20. Mugal, C. F., Weber, C. C. & Ellegren, H. *Bioessays* **37**, 1317–1326 (2015).
21. Lassalle, F. et al. *PLoS. Genet.* **11**, e1004941 (2015).
22. Lynch, M. *Proc. Natl Acad. Sci. USA* **109**, 18851–18856 (2012).
23. Long, H. et al. *Genome Biol. Evol.* **8**, 3815–3821 (2016).
24. McKenna, A. et al. *Genome Res.* **20**, 1297–1303 (2010).
25. DePristo, M. A. et al. *Nat. Genet.* **43**, 491–498 (2011).
26. Van der Auwera, G. A. et al. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–33 (2013).
27. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. *Brief. Bioinform.* **14**, 178–192 (2013).

Acknowledgements

Support was provided by the Multidisciplinary University Research Initiative awards W911NF-09-1-0444 and W911NF-14-1-0411 from the US Army Research Office to M.L., National Institutes of Health awards R01-GM036827 and R35-GM122566 to M.L., National Natural Science Foundation of China 31741071 to H.L., R01-GM51986 and R35-GM122556 to Y.V.B., F32-GM083581 to D.T.K. and National Science Foundation grant DOB 1442246 to J.T.L. We thank T. G. Doak, P. Keightley, K. Morris, R. Ness, I. Ruiz-Trillo, S. Simpson, W. K. Thomas, A. Uchimura and Z. Ye for providing strains and/or technical help in data acquisition. We thank L. Duret for helpful comments.

Author contributions

H.L., W.S., and M.L. conceived and designed the study, performed the data analyses and wrote the manuscript. All authors contributed to data collection and provided input to the manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-017-0425-y.

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

Correspondence and requests for materials should be addressed to M.L.

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

nature research

Corresponding author(s): Michael Lynch

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

We used all available 37 whole-genome mutation datasets, which include all high-quality/published data of this type. This gives sufficient statistical power for all correlation/regression analyses.

2. Data exclusions

Describe any data exclusions.

N/A

3. Replication

Describe whether the experimental findings were reliably reproduced.

Based on previous similar studies, we used ~50 replicates for each mutation accumulation experiment (12 experiments in total).

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

This study basically includes organisms with whole-genome mutation datasets available and sufficient number of accumulated mutations.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	R, Perl, Python, SigmaPlot

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No restriction.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	The only eukaryotic cell line--the Ichthyosporean Sphaeroforma arctica, was requested from Inaki Ruiz-Trillo's lab in Spain.
b. Describe the method of cell line authentication used.	Genome sequence and 18S rRNA
c. Report whether the cell lines were tested for mycoplasma contamination.	N/A
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	N/A