

MATTERS ARISING

<https://doi.org/10.1038/s41467-018-07950-z>

OPEN

Caution in inferring viral strategies from abundance correlations in marine metagenomes

Hend Alrasheed¹, Rong Jin¹ & Joshua S. Weitz ^{1,2}

ARISING FROM F.H. Coutinho et al. *Nature Communications* <https://doi.org/10.1038/ncomms15955> (2017).

Coutinho et al.¹ reported metagenomics-derived evidence in support of the ‘Piggyback-the-Winner’ (PtW) hypothesis that lysogeny prevalence increases at high microbial abundances. Coutinho et al.¹ did not directly estimate lysogenic prevalence, but instead, found that the ratio of virus-to-microbial host abundances decreased as microbial cell abundances increased. This pattern represents potential (albeit indirect) evidence in support of PtW. Here, we show that the bulk of these reported abundance relationships are likely spurious. Instead, we find absence of evidence for positive, sublinear correlations between virus and microbial abundances as estimated in dozens of putative virus-microbe pairs identified by Coutinho et al.¹ The absence of correlations between virus and microbial abundances is a counter-indicator for PtW. Altogether, our re-analysis suggests the need for caution in using correlation-based inference to identify viral strategies from metagenomics-derived abundance relationships.

To begin, consider the work of Coutinho et al.¹, who developed a metagenomics-based approach to characterize the diversity, ecology, host-associations, and strategies of marine phage. In doing so, they introduced a “new method for host prediction based on co-occurrence associations”, in which “virus–virus abundance associations were used for host affiliation”¹. As a result, Coutinho et al.¹ claimed that observed abundance relationships amongst phage and bacterial hosts in a range of marine habitats are consistent with the recently introduced mechanism of PtW².

The hypothesis underlying PtW is that viruses have increased lysogenic prevalence (and decreased lytic activity) with increasing microbial abundances. This hypothesis is meant to provide a mechanistic basis for empirical findings that total virus abundances increase with total microbial abundances even as the number of viruses per microbe decreases as microbial abundances increase. This pattern is found across marine, freshwater and other environmental systems (see Knowles et al.², Wigginton et al.³, and Parikka et al.⁴), with similar patterns found in predator–prey relationships⁵.

Sublinear (or less than proportional⁶) increases in virus abundances with microbial abundances may arise from multiple

governing mechanisms. These mechanisms include PtW, whose underlying mathematical model predicts that viral abundances increase with increasing microbial abundances, albeit sublinearly (see Fig. 1b of ref. ²). Additional mechanisms that could explain sublinear increases include variation in life history traits in antagonistic virus–microbe dynamics⁷ or trade-offs in Kill-the-Winner models⁸. As a consequence, the value of these patterns as exclusive indicators of any particular mechanism is disputed (see exchange of Weitz et al.⁷ and Knowles and Rohwer⁹, as well as the follow-up work of Knowles et al.¹⁰). Nonetheless, the possibility of using metagenomics-based methods to infer virus–host pairs and their abundance relationships could provide insights into viral strategies and their consequences in marine systems.

Here, we focus on the empirical findings of Coutinho et al.¹ and ask: do the abundance relationships exhibit robust evidence for sublinear increases in virus abundances with microbial abundances? Coutinho et al.¹ used multiple approaches, including virus–virus abundance associations, to link viruses and their putative hosts. We use the term “abundances” as a proxy for the metagenomics-inferred relative densities of viruses and host types reported by Coutinho et al.¹, consistent with their implementation. Once they estimated abundances, Coutinho et al.¹ quantified the relationship between the ratio of virus-to-host abundances vs. host abundances given putative pairs at both the genus and phylum levels. For example, let y be the log-transformed virus abundance and x be the log-transformed host abundance of an identified pair. If y increased sublinearly with x , then one would expect that $y \sim x^\alpha$ where $0 < \alpha < 1$. The inequality $\alpha > 0$ implies that virus abundances increase with microbial abundances and the inequality $\alpha < 1$ implies that the increase is sublinear.

It is also possible to evaluate ratio-based fits, i.e., quantifying the relationship between y/x and x . In that case, we expect $y/x \sim x^\beta$ where $\beta = \alpha - 1$. Hence, sublinear power-law relationships between y and x should lead to power-law relationships between virus-microbe ratios and microbial abundances with negative slopes between $-1 < \beta < 0$. Coutinho et al.¹ examined relationships between y/x vs. x , rather than directly examining y vs. x . If y is unrelated to x then one would expect best-fit curves between y/x vs. x to be statistically equivalent to fitting $1/x$ vs. x , thereby

¹School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA. ²School of Physics, Georgia Institute of Technology, Atlanta, GA 30332, USA. Correspondence and requests for materials should be addressed to J.S.W. (email: jswertz@gatech.edu)

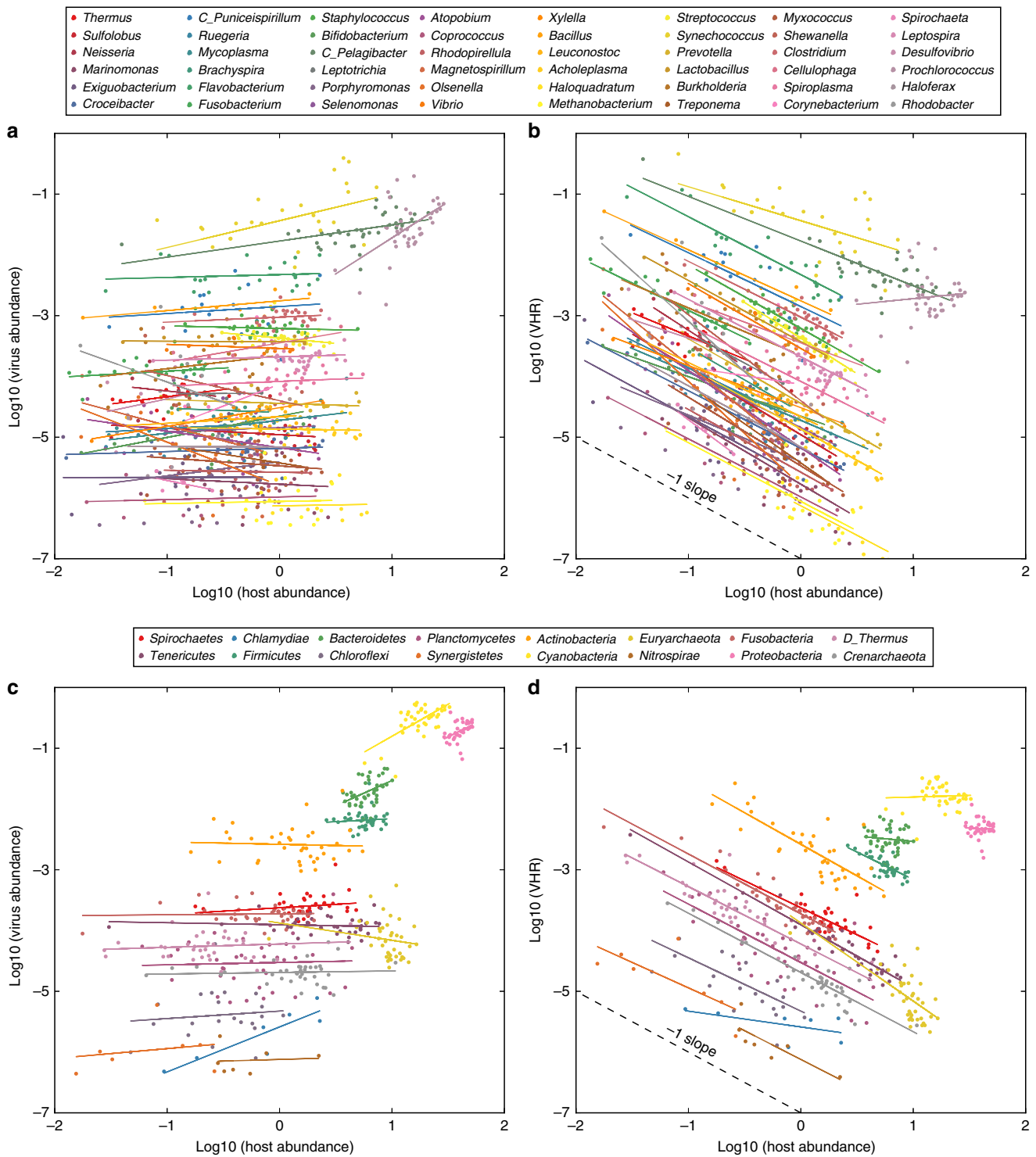


Fig. 1 Relationships between virus abundances, virus–host ratios (VHRs), and host abundances. **a** Genus-level relationships, virus vs. host abundances. **b** Genus-level relationships, VHRs vs. host abundances. **c** Phylum-level relationships, virus vs. host abundances. **d** Phylum-level relationships, VHRs vs. host abundances. Panels **b** and **d** are replotted from ref. ¹. The absence of a relationship between original, untransformed virus and host abundances (see panels **a** and **c**) should appear as a -1 slope when comparing VHR to host abundances (see panels **b** and **d**). A -1 slope is shown as a guide to readers in panels **b** and **d**

yielding a slope of $\beta = -1$ on a log–log plot. This relationship is an example of spurious self-correlation, in which inferences are derived based on correlating x against itself, in the absence of supporting evidence that y is correlated with x (see refs. ^{11,12}). Upon initial inspection, many reported slopes in Coutinho et al.¹ appear to be close to -1 (see re-plot of data in Fig. 1b, d).

This observation forms the basis for the present analysis. If slopes of y/x vs. x are in fact indistinguishable from -1 , then it would seem to indicate that there is not evidence that virus abundances, y , increase with microbial abundances, x . Moreover, we should not conclude that the power-law exponent α is greater than zero. Initial inspection suggests that there is a systematic

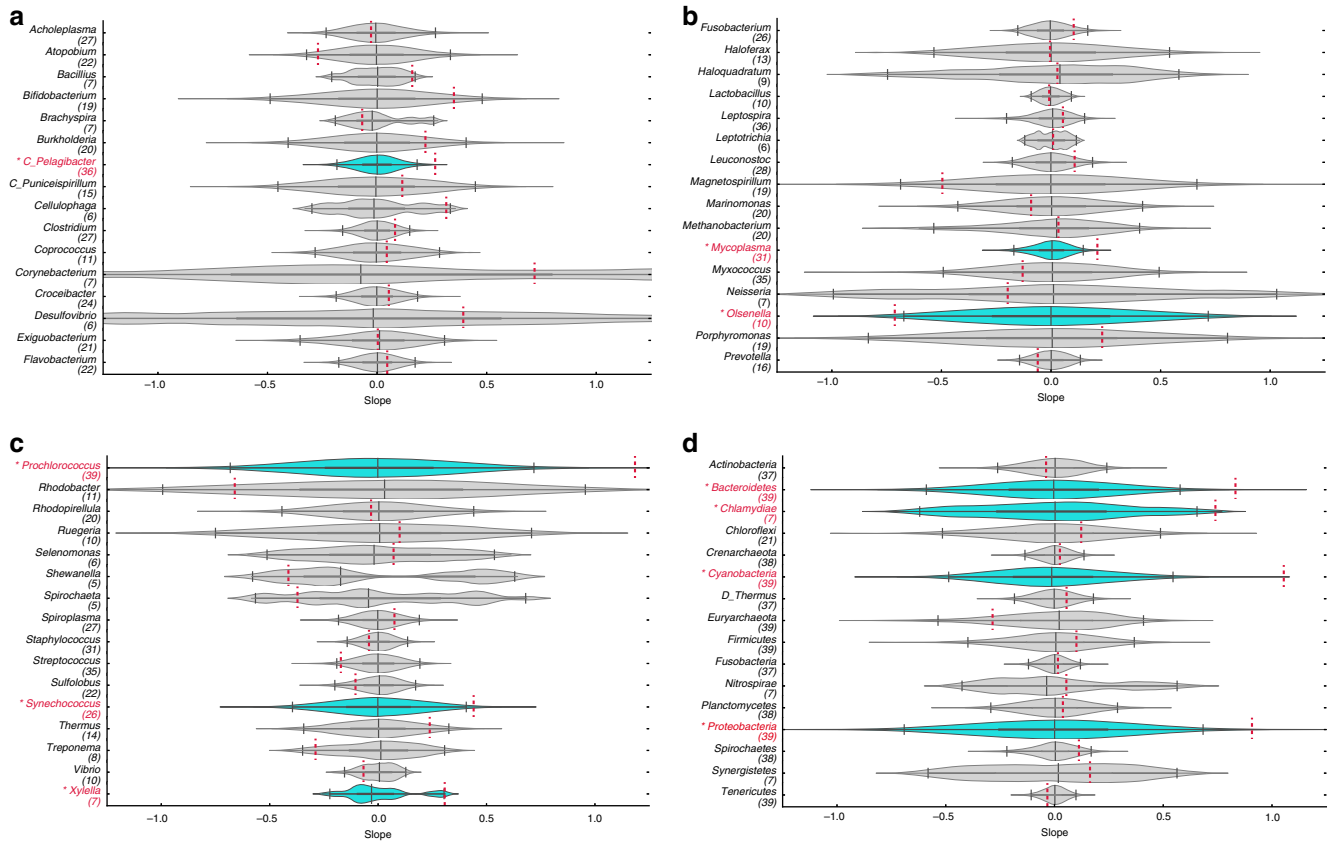


Fig. 2 Violin plots of the 95% randomized confidence intervals (CIs) of the slope for relationships between virus and host abundances. **a–c** CIs in genus datasets. **d** CIs in phylum datasets. In each panel, red dashed lines indicate the slope of the original sample. Dark vertical lines indicate the 2.5, 50, and 97.5 percentiles. Datasets with slopes with $p \geq 0.05$ are shown in gray violin plots and datasets with slopes with $p < 0.05$ are shown in blue violin plots. The number of data points for each dataset is indicated in the parenthesis below the genus/phylum name

absence of evidence for a relationship between y and x (see Fig. 1a, c). A lack of positive correlation would be surprising and not supportive of PtW (or the empirical literature) irrespective of whether or not the virus-to-microbe ratio decreased as microbes increased. In essence, by focusing on the inequality $\beta < 0$, it appears Coutinho et al.¹ did not fully consider evidence that $\alpha > 0$ (in a statistically significant sense).

To investigate this further, we evaluated the statistical relationship between y and x using a permutation test. The permutation test takes each host–virus pair and then permutes the virus abundances randomly, while maintaining the same host abundances. After each permutation, we recalculated the correlation between virus and host abundances across metagenomes. Such permutations should, in principle, eliminate latent correlations between the two variables. Yet, by chance, we might find some correlations even after permutation. Hence, this test enables a quantitative answer to the question: how often should there be slopes with at least as large a magnitude as observed in the data in the event that there were no underlying relationship between virus and host abundances? We can apply similar permutation methods for nonparametric slopes. In practice, we implemented a two-tailed randomized permutation test to generate a sampling distribution of expected slopes with mean 0.

We find that only 6 genus datasets (out of 48) and 4 phylum datasets (out of 16) have significant correlations at the $p = 0.05$ level (see Fig. 2). A strict Bonferroni correction suggests a criterion for significance of $p = 0.00078$ given the 64 (48 + 16) total comparisons. Hence, the threshold of $p = 0.05$ is permissive. Whether using $p = 0.05$ or $p = 0.01$ for product–moment or rank-based correlations, our results suggest that nearly all of the

relationships reported in Coutinho et al.¹, do not show evidence of a relationship between virus and host abundance and therefore do not provide even indirect support for PtW (see Supplementary Tables 1–4 for confidence intervals for all relationships examined). Critically, our findings are consistent with calculations of the significance of virus–host correlations reported in Supplementary Tables 3 and 4 of Coutinho et al.¹, who nonetheless claimed that their “findings corroborate the recently proposed Piggyback-the-Winner theory.”

Given this reanalysis, we conclude that the abundance relationships derived from marine metagenomics datasets in Coutinho et al.¹ do not have robust evidence in support of PtW. To the contrary, nearly the entire dataset of putative relationships between viruses and hosts do not have evidence that virus abundances are related to host abundances. It is this lack of a relationship (i.e., $\alpha \approx 0$) that explains why the observation that the ratio of viruses to hosts decreases with increasing host abundances (i.e., $\beta \approx -1$) was incorrectly interpreted in Coutinho et al.¹. Such a decrease can occur in the absence of a relationship between virus abundances and host abundances. This has been termed spurious self-correlation^{11,12}.

Our reanalysis raises a number of questions. The absence of significant abundance relationships between viruses and hosts may be a consequence of analyzing correlations based on relative abundances, which can lead to spurious findings¹³. The absence of a relationship may also indicate that the method for host prediction developed by Coutinho et al.¹ fails to identify genuine interactions. In that event, the absence of a correlation may reflect the absence of an underlying interaction between putative virus–host pairs. However, correlation does not imply causation,

just as the absence of correlation does not imply the absence of causation. Multiple studies have reported that correlation in microbe–microbe dynamics and virus–microbe dynamics may be a poor indicator of interaction^{14–16}.

We recognize that correlation-based inference of virus–host interactions was one of multiple methods used by Coutinho et al.¹ to identify virus–host pairs. If the methods of Coutinho et al.¹ accurately identify virus–host pairs, then our re-examination suggests that abundance relationships between viruses and microbes may differ starkly when considered amongst lineages¹ vs. when given total abundance data^{2–4}. If robust, the absence of a significant relationship between virus and microbial abundances given lineage-specific interactions may result from other eco-evolutionary drivers, e.g., “cryptic dynamics” in which rapid evolution masks latent abundance relationships¹⁷.

In summary, we cannot say definitively whether the identified virus–microbes associations in Coutinho et al.¹ are functionally relevant. However, we can conclude that the abundance relationships inferred from metagenomes do not provide robust, indirect support for PtW (or other mechanistic hypotheses) that predict sublinear increases in viral abundances with microbial abundances. Moving forward, we hope that the combined use of new in situ technologies and principled, in silico analyses can help advance the identification of likely virus–host interactions, viral strategies, and population-level consequences of viral infection.

Methods

Datasets. The datasets of Coutinho et al.¹ were partitioned based on whether interactions were identified at the genus or phylum scale. In the original dataset, there were 17 and 93 virus–host datasets at the phylum scale and genus scale respectively. Each dataset has 39 distinct sampling sites. However, those sites with zero virus abundance or zero host abundance were excluded. In addition, we did not include any datasets for which there were not at least five sample sites with both non-zero levels of microbes and viruses. As a result, the datasets analyzed included 16 of 17 original virus–host pairs at the phylum level and 48 of 93 original virus–host pairs at the genus level.

Statistical analysis. We implemented permutation tests to assess measured correlations between virus and host abundances. For each host dataset we generated 10^4 randomized samples of size n_i (the number of sample sites for virus–host pair i). In each sample, we permuted the virus abundances without replacement while holding the host abundances fixed. Distributions of measured correlations were compared to the original samples using product–moment correlation (main text and Supplementary Information) and Spearman rank correlation (see Supplementary Information).

Data availability

The original data is from Coutinho et al.¹. All statistical analysis code and results are available via an open access link at <https://doi.org/10.5281/zenodo.1478122>.

Received: 3 May 2018 Accepted: 6 December 2018

Published online: 30 January 2019

References

- Coutinho, F. H. et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat. Commun.* **8**, 15955 EP (2017).
- Knowles, B. et al. Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).
- Wigington, C. H. et al. Re-examination of the relationship between marine virus and microbial cell abundances. *Nat. Microbiol.* **1**, 15024 (2016).
- Parikka, K., Le Romancer, M., Wauters, N. & Jacquet, S. Deciphering the virus-to-prokaryote ratio (VPR): insights into virus–host relationships in a variety of ecosystems. *Biol. Rev.* **92**, 1081–1100 (2016).
- Hatton, I. A. et al. The predator–prey power law: biomass scaling across terrestrial and aquatic biomes. *Science* **349**, aac6284 (2015).
- Cael, B. B., Carlson, M. C. G., Follett, C. L. & Follows, M. J. Marine virus-like particles and microbes: a linear interpretation. *Front. Microbiol.* **9**, 358 (2018).
- Weitz, J. S., Beckett, S. J., Brum, J. R., Cael, B. & Dushoff, J. Lysis, lysogeny, and virus–microbe ratios. *Nature* **549**, E1–E3 (2017).
- Thingstad, T. F. & Bratbak, G. Microbial oceanography: viral strategies at sea. *Nature* **531**, 454–455 (2016).
- Knowles, B. & Rohwer, F. Knowles & Rohwer reply. *Nature* **549**, E3–E4 (2017).
- Knowles, B. et al. Variability and host density independence in inductions-based estimates of environmental lysogeny. *Nat. Microbiol.* **2**, 17064 (2017).
- Kenney, B. Beware of spurious self-correlations! *Water Resour. Res.* **18**, 1041–1048 (1982).
- Jackson, D. & Somers, K. The spectre of “spurious” correlations. *Oecologia* **86**, 147–151 (1991).
- Kurtz, Z. D. et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, 1–25 (2015).
- Stein, R. R. et al. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol.* **9**, e1003388 (2013).
- Fisher, C. K. & Mehta, P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS ONE* **9**, e102451 (2014).
- Coenen, A. R. & Weitz, J. S. Limitations of correlation-based inference in complex virus–microbe communities. *mSystems* e00084-18 (2018).
- Cortez, M. H. & Ellner, S. P. Understanding rapid evolution in predator–prey interactions using the theory of fast-slow dynamical systems. *Am. Nat.* **176**, E109–E127 (2010).

Acknowledgments

The authors thank Bas Dutilh and Matt Sullivan for helpful comments and discussions. This work was supported by the Simons Foundation (SCOPE award ID 329108, J.S.W.).

Author contributions

H.A., R.J., and J.S.W. designed simulations, performed simulations, and analyzed data. J.S.W. designed the study and wrote the manuscript with contributions from H.A. and R.J.

Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-018-07950-z>.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

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

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) 2019