AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

OBSERVATION

Predicting Parallelism and Quantifying Divergence in Microbial Evolution Experiments

William R. Shoemaker,[a]* Jay T. Lennon[a]

[a]Department of Biology, Indiana University, Bloomington, Indiana, USA

ABSTRACT The degree to which independent populations subjected to identical environmental conditions evolve in similar ways is a fundamental question in evolution. To address this question, microbial populations are often experimentally passaged in a given environment and sequenced to examine the tendency for similar mutations to repeatedly arise. However, there remains the need to develop an appropriate statistical framework to identify genes that acquired more mutations in one environment than in another (i.e., divergent evolution), genes that serve as genetic candidates of adaptation. Here, we develop a mathematical model to evaluate evolutionary outcomes among replicate populations in the same environment (i.e., parallel evolution), which can then be used to identify genes that contribute to divergent evolution. Applying this approach to data sets from evolve-and-resequence experiments, we found that the distribution of mutation counts among genes can be predicted as an ensemble of independent Poisson random variables with zero free parameters. Building on this result, we propose that the degree of divergent evolution at a given gene between populations from two different environments can be modeled as the difference between two Poisson random variables, known as the Skellam distribution. We then propose and apply a statistical test to identify specific genes that contribute to divergent evolution. By focusing on predicting patterns among replicate populations in a given environment, we are able to identify an appropriate test for divergence between environments that is grounded in first principles.

IMPORTANCE There is currently no universally accepted framework for identifying genes that contribute to molecular divergence between microbial populations in different environments. To address this absence, we developed a null model to describe the distribution of mutation counts among genes. We find that divergent evolution within a given gene can be modeled as the absolute difference in the total number of mutations observed between two environments. This quantity is effectively captured by a probability distribution known as the Skellam distribution, providing an appropriate statistical test for researchers seeking to identify the set of genes that contribute to divergent evolution in microbial evolution experiments.

KEYWORDS experimental evolution, microbial evolution, evolution, parallel evolution, adaptation

Biologists have long been fascinated by the degree to which evolution is repeatable (1). Independently evolving microbial populations frequently evolve similar genotypes and phenotypes, a phenomenon known as parallel evolution (2, 3). Through the rise of evolve-and-resequence experiments as high-throughput screens for adaptation, researchers can now identify recurrent mutations across replicate populations to pare down the vast number of potentially adaptive mutations into those that putatively confer the largest fitness benefits (4, 5). Furthermore, evolve-and-resequence experiments have revealed that the outcomes of evolution are often conditional on ancestral genotype (6–10) as well as the environment in which the microbial populations were maintained (11–16), a phenomenon known as divergent evolution.

Editor Michael J. Imperiale, University of Michigan—Ann Arbor

Address correspondence to William R. Shoemaker, williamrshoemaker@gmail.com.

*Present address: William R. Shoemaker, Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, California, USA.

The authors declare no conflict of interest.

Received 20 August 2021
Accepted 28 January 2022
Published 9 February 2022

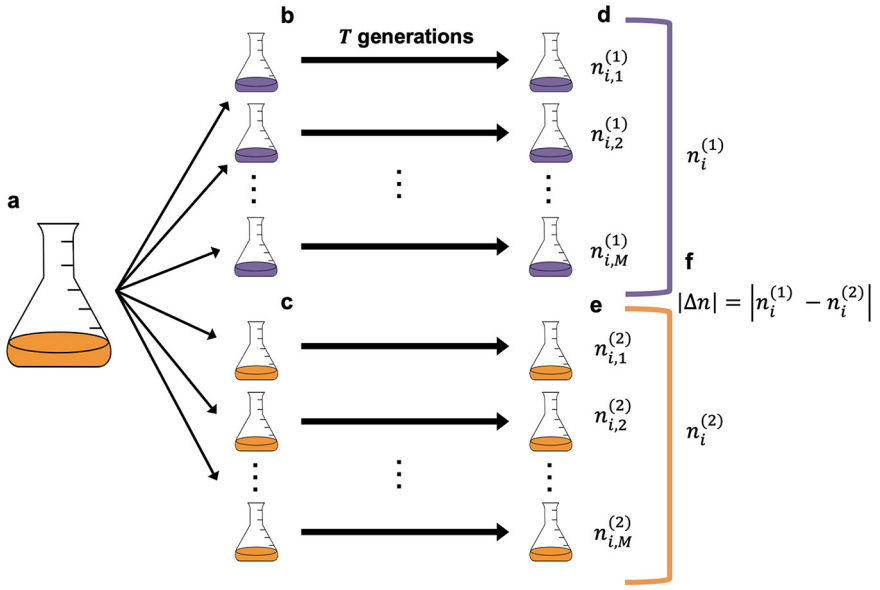

FIG 1 (a) A typical evolve-and-resequence experiment is performed by splitting a culture that has been grown from a single colony, inoculating cells into replicate flasks constituting one or more environmental conditions (e.g., purple or orange), and propagating the population over time by periodically transferring cells into new flasks with fresh medium. (b and c) After a given number of generations has elapsed, replicate populations are often sequenced, allowing the number of *de novo* mutations at a given gene to be calculated. (d to f) The degree of parallel evolution within each environment is quantified by taking the sum of mutation counts across replicate populations for a given gene (d and e), while the degree of divergent evolution is quantified by taking the absolute difference in mutation counts between environments ($|\Delta n|$) (f).

Despite the potential power of evolve-and-resequence experiments, statistical frameworks to quantify the repeatability of evolution are lacking. In recent years, models that coarse-grain over molecular details have been remarkably successful in identifying general evolutionary principles (17). This approach, and the underlying motivations to develop straightforward interpretations of biological phenomena, raises the question of whether there are intuitive ways in which the contributors to divergent evolution can be identified. To address this task, we first determined the extent to which patterns of parallel evolution at the gene level can be predicted using a statistical model containing zero free parameters with publicly available data. Building on these results, we formulated and tested an interpretable null model of divergent evolution at the gene level. In both cases, we use data from published experiments with bacteria, but in principle, the statistical methods can be applied to populations of archaea, microeukaryotes, and viruses.

Predicting genetic parallelism among replicate populations. The task of identifying genes that contribute to divergent evolution can be viewed as the equivalent of identifying genes that undergo a greater degree of parallel evolution in one environment than in another environment (Fig. 1). This observation suggests that it is necessary to first identify an appropriate model of parallel evolution within a single environment in order to develop a null model of divergence. Given that the per-generation probability of acquiring a mutation at a given gene is low and the number of generations is large, it is reasonable to assume that a given gene acquires mutations as a Poisson process. We can model the sampling distribution of this process as the probability of observing $n_{i,j}$ mutations in the ith gene within a population that acquired a total of $n_{\text{tot},j}$ mutations:

$$P(n_{i,j}|n_{\text{tot},j}) = \binom{n_{\text{tot},j}}{n_{i,j}} \left(\frac{n_{i,j}}{\sum_i n_{i,j}}\right)^{n_{i,j}} \left(\frac{\sum_{k\neq i} n_{k,j}}{\sum_i n_{i,j}}\right)^{n_{\text{tot},j}-n_{i,j}} \quad (1)$$

We can then determine whether we can predict statistical patterns from empirical data using equation 1. Given that mutation count data from evolve-and-resequence

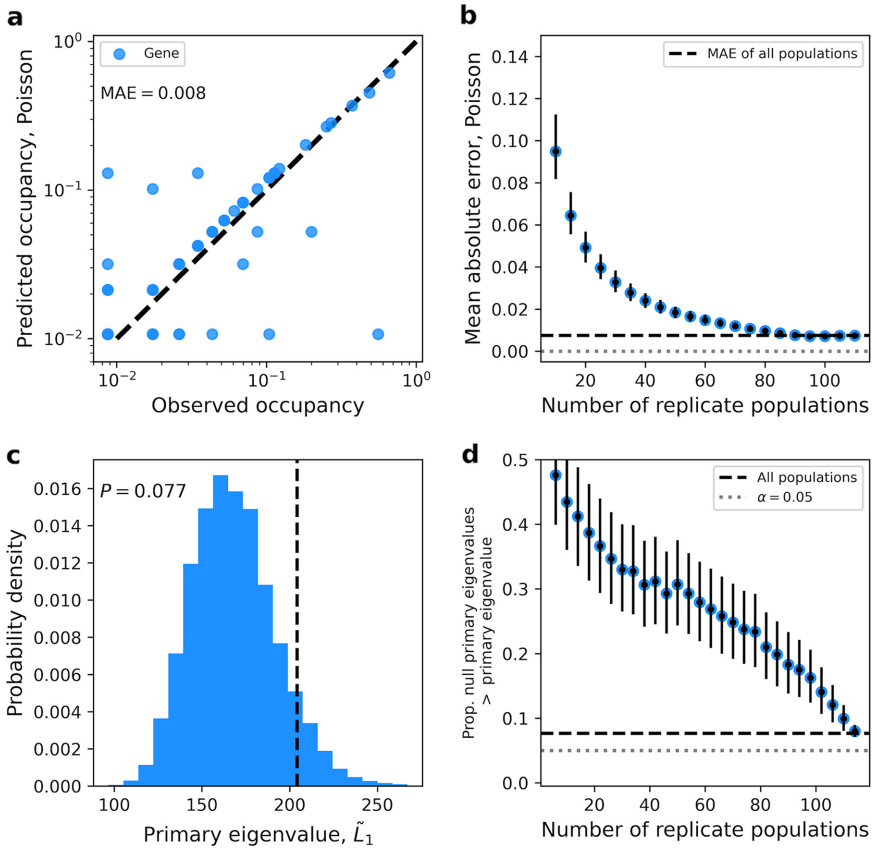

FIG 2 (a) Using the Poisson distribution, we were able to predict the occupancy of nonsynonymous mutations for a given gene among 115 replicate *E. coli* populations. (b) Using the same data set, we were able to subsample replicate populations to examine how the level of error in our prediction decreased as the number of replicate populations increased. (c) The degree of covariance between genes is summarized by the primary eigenvalue of the gene-by-population matrix of mutation counts (dashed black line). By generating null count matrices, we simulated a null distribution of primary eigenvalues to calculate the *P* value for the observed degree of covariance. (d) Similar to the analysis in panel c, we examined how the ability to detect covariance changes as the number of replicate populations increases by calculating the fraction of observed primary eigenvalues greater than the null.

experiments are often sparse (i.e., zeros comprise a large proportion of the observations), it is natural to calculate the proportion of populations that have at least one mutation in a given gene (i.e., occupancy, o_i [18]) and compare our empirical estimate to an expected value by averaging over *M* replicate populations:

$$o_i = 1 - \frac{1}{M}\sum_{j}^{M} P\left(0|n_{\mathrm{tot},j}\right) = 1 - \frac{1}{M}\sum_{j}^{M}\left(\frac{\sum_{k\neq i} n_{k,j}}{\sum_{i} n_{i,j}}\right)^{n_{\mathrm{tot},j}} \tag{2}$$

To test this prediction, we calculated $\langle o_i \rangle$ from empirical data. Given that the number of genes (i.e., variables) is typically much larger than the number of mutations (i.e., observations) within a typical population in an evolve-and-resequence experiment, it is necessary to examine an experiment that maintained a large number of replicates. The *Escherichia coli* evolve-and-resequence data set from Tenaillon et al. contains 115 replicate populations that originated from a single genotype (i.e., CFU) and were passaged for 2,000 generations (11), a number that was, and still is, far larger than that of a typical evolve-and-resequence experiment. We found that our model does a reasonable job capturing the observed occupancy of nonsynonymous mutations (Fig. 2a), with a mean absolute error (MAE) of ~0.008. The success of the Poisson model is even more apparent when it is compared to a reasonable alternative model where each lineage

can acquire a maximum of one mutation in a given gene (see Text S1, equation S2, and Fig. S1 in the supplemental material). However, while the MAE decreased with an increasing number of replicate populations, it ultimately saturated (Fig. 2b). The fact that it does not reach zero suggests that features not incorporated into our model, such as nonindependence among genes, may be necessary to fully explain the distribution of mutation counts.

To determine whether nonindependence among genes was necessary to incorporate in our model, we tested whether we could detect signals of covariance in our data. Because the number of genes that acquired mutations in an experiment can be in the hundreds, and mutation count data are sparse, attempting to estimate individual covariances for all pairs of genes would be unreasonable. Instead, we estimated a global signature of covariance and compared it to an appropriate null distribution (see "Predicting and quantifying parallelism," below). While the global signal of covariance increased with the number of replicate populations, it was weak for values typical of most evolution experiments (5 to 20 populations) (Fig. 2c and d) and was only borderline significant when all 115 replicate populations were included ($P = 0.072$). This result suggests that we can proceed with the development of a null model of divergent evolution without the incorporation of covariance between genes, instead modeling the degree of parallel evolution for a given gene as an independent random variable.

Identifying genes that contribute to divergent evolution between environments. The success of the multivariate Poisson model in describing the distribution of mutation counts within a given environment along with the overall weak signals of covariance provide justification for modeling the distribution of mutation counts among genes as an ensemble of effectively independent variables. We can then model divergent evolution at a given gene as the difference between two independent Poisson rates. In terms of mutation counts, we can identify the meaningful variable as the absolute difference in mutation counts between two environments for a given gene ($|\Delta n_i| = |n_i^{(1)} - n_i^{(2)}|$). The distribution of $|\Delta n|$ has been previously derived and is known as the Skellam distribution (19). Starting with the null Poisson rates for each treatment ($\lambda_1 = n_{\text{tot}}^{(1)}/N_{\text{genes}}$; $\lambda_2 = n_{\text{tot}}^{(2)}/N_{\text{genes}}$), we define the probability mass function of the absolute value of $|\Delta n|$ as

$$\Pr[|\Delta n| \,|\lambda_1,\, \lambda_2] = \begin{cases} e^{-(\lambda_1+\lambda_2)}\left[\left(\frac{\lambda_1}{\lambda_2}\right)^{\frac{|\Delta n|}{2}} I_{\Delta n}\left(2\sqrt{\lambda_1\lambda_2}\right) + \left(\frac{\lambda_2}{\lambda_1}\right)^{\frac{|\Delta n|}{2}} I_{-\Delta n}\left(2\sqrt{\lambda_1\lambda_2}\right)\right] & \text{if } |\Delta n| > 0 \\ e^{-(\lambda_1+\lambda_2)} I_0\left(2\sqrt{\lambda_1\lambda_2}\right) & \text{if } |\Delta n| = 0 \end{cases} \tag{3}$$

Where $I_{\Delta n}(\cdot)$ is a modified Bessel function of the first kind. Building on a previous approach developed to identify contributors of parallel evolution (20), we can define the P value as

$$P_i = \sum_{|\Delta n| \geq |\Delta n_i|} \Pr[|\Delta n| \,|\lambda_1,\, \lambda_2] \tag{4}$$

To reduce the number of tests, we can calculate P values only for $|\Delta n| \geq n_{\text{min}}$, where the expected number of genes with $|\Delta n| \geq n_{\text{min}}$ and $P_i \leq P$ is

$$\overline{N}(P) \approx \sum_{i=1}^{N_{\text{genes}}} \sum_{|\Delta n| = n_{\text{min}}}^{\infty} \theta(P - P_i(|\Delta n|)) \cdot \Pr[|\Delta n| \,|\lambda_1,\, \lambda_2] \tag{5}$$

where $\theta(\cdot)$ is the Heaviside step function. We can then compare this number to the observed number of genes, $N(P)$, defining a critical P value (P^*) for a given false discovery rate (FDR), α, as

$$\frac{\overline{N}(P^*)}{N(P^*)} \leq \alpha \tag{6}$$

To apply this approach, it was necessary to identify an evolve-and-resequence experiment that maintained at least two treatments. We identified an appropriate experiment where six replicate populations of the bacterium *Burkholderia cenocepacia* were propagated for ~600 generations in four environments, allowing us to identify the set of genes that were consistently enriched for nonsynonymous mutations within a given treatment across all pairwise treatment comparisons (12) (Table S1). Our results largely agree with the conclusions of the original study: virtually all the genes that were significantly enriched within a single treatment in the original study were also identified as contributors to environment-specific adaptation (12).

Concluding remarks. We investigated the distribution of mutation counts in bacterial evolve-and-resequence experiments. We found that a Poisson model containing zero free parameters sufficiently explained the distribution of mutation counts across genes. This result suggests that parallel evolution among replicate populations in evolve-and-resequence experiments can be quantitatively predicted without the use of models that require statistical fits (e.g., linear regression). We then developed an intuitive null model for identifying genes that contributed to the genetic differences that accrued between bacterial populations that evolved in different environments (i.e., divergent evolution). Using this result, the difference in the numbers of mutations within a given gene between treatments ($|\Delta n|$) can be modeled as a difference in Poisson rates between treatments (i.e., the Skellam distribution).

Our approach should be robust to documenting parallel and divergent evolution in evolve-and-resequence experiments with diverse microbial taxa. While we focused on bacterial case studies owing to features related to experimental design, there is no reason why the framework cannot be applied to archaea, microeukaryotes, and viruses. One biological feature that may require additional consideration is the existence of sexual recombination in eukaryotes. However, this is unlikely to substantially alter our results, as recombination breaks the physical linkage between mutations, which subsequently reduces the magnitude of covariance between a given pair of genes. In addition, we note that the facilitation of recombination is the principal effect of sexual reproduction on the molecular evolutionary dynamics of a population, an evolutionary force that often occurs in bacteria. While we did not determine the extent to which the molecular details of recombination affect the accuracy of the Poisson model, and recombination rates are difficult to infer in bacterial evolve-and-resequence experiments, we note that there was evidence of homologous recombination in the experimental data that we examined (11), suggesting that the presence or absence of recombination alone is insufficient to substantially affect our predictions.

Data. To determine the degree to which we can predict statistical patterns of parallel evolution, we used a publicly available data set of one of the largest microbial evolve-and-resequence experiments. In this experiment, 115 replicate populations of *Escherichia coli* were serially transferred for 2,000 generations at 42.2°C (11). A single colony was isolated from each replicate population and sequenced at the end of the experiment.

To apply our divergence test, we used a publicly available data set from a factorially designed experiment where the bacterium *Burkholderia cenocepacia* was propagated for ~600 generations at 37°C in a roller drum. Specifically, replicate populations ($n = 6$) were grown in either low- or high-carbon medium in the presence or absence of a bead that was used to promote biofilm or planktonic growth, respectively.

Predicting and quantifying parallelism. To test for a global signal of covariance between genes, we merged all nonsynonymous mutations from all replicate populations into a population-by-gene-count matrix. To account for gene size as a covariate, we corrected the number of mutations by calculating the excess number of mutations (i.e., multiplicity), $m_{i,j} = n_{i,j} \cdot \overline{L}/L_i$, where L_i is the number of nonsynonymous sites in the ith gene and $\overline{L}$ is the mean of all genes in the genome (20). To determine whether

covariance could be reliably detected at a given level of replication, we estimated the largest normalized eigenvalue over the set of eigenvalues ($\{E_1\}$) (21, 22), defined as

$$\tilde{e}_1 = \frac{e_1 - \mu(M, N_{\text{genes}})}{\sigma(M, N_{\text{genes}})} \tag{7}$$

where e_1 is normalized as $e_1 = ME_1 / \sum_{m=1}^{M} E_m$ to sum to M, E_1 is the largest eigenvalue, and

$$\mu(M, N_{\text{genes}}) = \frac{\left(\sqrt{N_{\text{genes}} - 1} + \sqrt{M}\right)^2}{N_{\text{genes}}} \tag{8}$$

$$\sigma(M, N_{\text{genes}}) = \frac{\sqrt{N_{\text{genes}} - 1} + \sqrt{M}}{N_{\text{genes}}} \left(\frac{1}{\sqrt{N_{\text{genes}} - 1}} + \frac{1}{\sqrt{M}}\right)^{\frac{1}{3}} \tag{9}$$

As M, $N_{\text{genes}} \to \infty$ and $N_{\text{genes}}/M \to \gamma \geq 1$, $\tilde{e}_1$ tends toward the Tracy-Widom distribution (22, 23), although these limits can be relaxed (21, 24). A null distribution of $\tilde{e}_1$ was obtained by randomizing combinations of mutation counts constrained on the total number of mutations acquired within each gene across treatments and the number of mutations acquired within each treatment. Randomization was performed using a Python implementation of the ASA159 algorithm (25, 26).

Data availability. Instructions and code to reproduce our analyses are on GitHub (https://github.com/LennonLab/ParEvol). All processed data are available on Zenodo (https://zenodo.org/record/3779341).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, DOCX file, 0.01 MB.

FIG S1, TIF file, 2.8 MB.

TABLE S1, TIF file, 0.6 MB.

ACKNOWLEDGMENTS

This work was supported by U.S. Army Research Office grant W911NF-14-1-0411 (J.T.L.), the National Science Foundation (DEB-1934554 and DBI-2022049 to J.T.L.), the National Aeronautics and Space Administration (80NSSC20K0618 to J.T.L.), and the Society for the Study of Evolution Graduate Research Excellent Grant Rosemary Grant Advanced Award (W.R.S.).

REFERENCES

1. Gould SJ. 1990. Wonderful life: the Burgess Shale and the nature of history. Norton & Co, New York, NY.
2. Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, Kingsley DM. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307:1928–1933. https://doi.org/10.1126/science.1107239.
3. Bolnick DI, Barrett RDH, Oke KB, Rennison DJ, Stuart YE. 2018. (Non)parallel evolution. Annu Rev Ecol Evol Syst 49:303–330. https://doi.org/10.1146/annurev-ecolsys-110617-062240.
4. Cooper VS. 2018. Experimental evolution as a high-throughput screen for genetic adaptations. mSphere 3:e00121-18. https://doi.org/10.1128/mSphere.00121-18.
5. McDonald MJ. 2019. Microbial experimental evolution—a proving ground for evolutionary theory and a tool for discovery. EMBO Rep 20:e46992. https://doi.org/10.15252/embr.201846992.
6. Vogwill T, Kojadinovic M, Furió V, MacLean RC. 2014. Testing the role of genetic background in parallel evolution using the comparative experimental evolution of antibiotic resistance. Mol Biol Evol 31:3314–3323. https://doi.org/10.1093/molbev/msu262.
7. Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. 2006. Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. Proc Natl Acad Sci U S A 103:9107–9112. https://doi.org/10.1073/pnas.0602917103.
8. Bailey SF, Guo Q, Bataillon T. 2018. Identifying drivers of parallel evolution: a regression model approach. Genome Biol Evol 10:2801–2812. https://doi.org/10.1093/gbe/evy210.
9. Bertels F, Leemann C, Metzner KJ, Regoes RR. 2019. Parallel evolution of HIV-1 in a long-term experiment. Mol Biol Evol 36:2400–2414. https://doi.org/10.1093/molbev/msz155.
10. Fisher KJ, Kryazhimskiy S, Lang GI. 2019. Detecting genetic interactions using parallel evolution in experimental populations. Philos Trans R Soc Lond B Biol Sci 374:20180237. https://doi.org/10.1098/rstb.2018.0237.
11. Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. 2012. The molecular diversity of adaptive convergence. Science 335:457–461. https://doi.org/10.1126/science.1212986.
12. Turner CB, Marshall CW, Cooper VS. 2018. Parallel genetic adaptation across environments differing in mode of growth or resource availability. Evol Lett 2:355–367. https://doi.org/10.1002/evl3.75.

13. Shoemaker WR, Jones SE, Muscarella ME, Behringer MG, Lehmkuhl BK, Lennon JT. 2021. Microbial population dynamics and evolutionary outcomes under extreme energy limitation. Proc Natl Acad Sci U S A 118: e2101691118. https://doi.org/10.1073/pnas.2101691118.
14. Shoemaker WR, Polezhaeva E, Givens KB, Lennon JT. 2021. Molecular evolutionary dynamics of energy limited microorganisms. Mol Biol Evol 38: 4532–4545. https://doi.org/10.1093/molbev/msab195.
15. Shoemaker WR, Polezhaeva E, Givens KB, Lennon JT. 2021. Seed banks alter the rate and direction of molecular evolution in Bacillus subtilis. bioRxiv https://doi.org/10.1101/2021.10.05.463161.
16. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, Wu GC, Wielgoss S, Cruveiller S, Médigue C, Schneider D, Lenski RE. 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature 536:165–170. https://doi.org/10.1038/nature18959.
17. Good BH, Hallatschek O. 2018. Effective models and the search for quantitative principles in microbial evolution. Curr Opin Microbiol 45:203–212. https://doi.org/10.1016/j.mib.2018.11.005.
18. Grilli J. 2020. Macroecological laws describe variation and diversity in microbial communities. Nat Commun 11:4743. https://doi.org/10.1038/s41467-020-18529-y.
19. Skellam JG. 1946. The frequency distribution of the difference between two Poisson variates belonging to different populations. J R Stat Soc Ser A 109:296. https://doi.org/10.2307/2981372.
20. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017. The dynamics of molecular evolution over 60,000 generations. Nature 551: 45–50. https://doi.org/10.1038/nature24287.
21. Tracy CA, Widom H. 1994. Level-spacing distributions and the Airy kernel. Commun Math Phys 159:151–174. https://doi.org/10.1007/BF02100489.
22. Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet 2:e190. https://doi.org/10.1371/journal.pgen.0020190.
23. Johnstone IM. 2001. On the distribution of the largest eigenvalue in principal components analysis. Ann Stat 29:295–327. https://doi.org/10.1214/aos/1009210544.
24. Soshnikov A. 2002. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. J Stat Phys 108: 1033–1056. https://doi.org/10.1023/A:1019739414239.
25. Baak M, Koopman R, Snoek H, Klous S. 2020. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. Comput Stat Data Anal 152:107043. https://doi.org/10.1016/j.csda.2020.107043.
26. Patefield WM. 1981. Algorithm AS159. An efficient method of generating R $\times$ C tables with given row and column totals. J R Stat Soc Ser C Appl Stat 30:91–97.