Predicting Parallelism and Quantifying Divergence in Experimental Evolution

William R. Shoemaker[1,2,*] **and Jay T. Lennon**[1]

[1]Department of Biology, Indiana University, Bloomington, IN, 47405, USA.
[2]Present affiliation: Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA.
[*]williamrshoemaker@gmail.com

ABSTRACT

ABSTRACT: The degree that the environment determines what genes contribute towards adaptation is a fundamental question in microbial evolution. Microbial populations are often experimentally passaged in different environments and sequenced in order to identify candidates for adaptation in a particular environment. However, there remains the need to develop an appropriate statistical framework to identify genes that acquired more mutations in one environment over the other (i.e., divergent evolution). Here we demonstrate how the evolutionary outcomes among replicate populations in the same environment, known as parallel evolution, can be leveraged to construct an intuitive statistical test for identifying the genes that contribute towards divergent evolution. To accomplish this task, we examined publicly available evolve-and-resequence experiment datasets and found that the distribution of mutation counts among genes can be predicted using an ensemble of independent Poisson random variables. 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 towards divergent evolution. IMPORTANCE: There is currently no existing framework that can be leveraged to identify genes that contribute towards divergent evolution in microbial evolution experiments. To correct for this absence, we investigated the distribution of mutation counts among genes in order to identify an appropriate null model. Our observations suggest that divergent evolution within a given gene can be modeled as the difference in the total number of mutations observed between two environments. This quantity is described 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 towards divergent evolution in evolution experiments.

Observation

Biologists have long been fascinated by the degree to which evolution is repeatable[1]. Independently evolving populations frequently evolve similar genotypes and phenotypes, a phenomenon known as parallel evolution[2,3]. Parallel evolution is particularly prevalent among microorganisms. The rise of evolve-and-resequence experiments as high-throughput screens for adaptation[4] has allowed researchers to identify recurrent mutations across replicate populations[4,5], paring down the vast number of potentially adaptive mutations into those that putatively confer the largest fitness benefits. Furthermore, evolve-and-resequence experiments have revealed that the outcomes of evolution are often conditional on the ancestral genotype of a microbial population[6–10] or the environment in which it was maintained[11–15], a phenomenon known as divergent evolution.

The ease in which evolve-and-resequence experiments can be performed comes with the drawback that there is comparatively little statistical direction on how the contributors of divergent evolution should be identified. In recent years, models that coarse-grain over molecular details have been remarkably successful in identifying general microbiological principles[16]. These models, and the underlying motivation to develop straightforward interpretations of biological phenomena, raises the question of whether there is an intuitive way in which contributors towards divergent evolution can be identified. To address this issue, we first determined the extent that we can predict patterns of parallel evolution at the gene level using a straightforward statistical model and publicly available data. Building on these results, we formulated and tested an interpretable model of divergent evolution at the gene level.

Predicting genetic parallelism among replicate populations

The task of identifying genes that contribute towards divergent evolution can be viewed as the equivalent of identifying genes that undergo a greater degree of parallel evolution in one environment relative to another (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 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 as

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

We can then determine whether we can predict statistical patterns from empirical data using Eq.1. Given that mutation count data from evolve-and-resequence experiments are often sparse, it is natural to calculate the proportion of populations that have at least one mutation in a given gene (i.e., *occupancy*, o_i[17]) and compare our empirical estimate to an expected value by averaging over M replicate populations

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

To test our prediction, we calculated $\langle o_i \rangle$ from Eq. 2 on nonsynonymous mutation count data from an evolve-and-resequence experiment with 115 replicate populations of *E. coli*[11]. We found that our model does a reasonable job capturing the observed occupancy (Fig. 2a) with a mean absolute error (MAE) of ~ 0.008. However, while 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 non-independence among genes may be necessary to fully explain the distribution of mutation counts.

To determine whether non-independence 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 number in the hundreds and mutation count data is sparse, attempting to estimate individual covariances would be unreasonable. Instead, we can estimate a global signature of covariance and compare it to a null distribution (Methods). While the global signal of covariance increased with the number of replicate populations, it was weak for values typical of most evolution experiments (5-20 populations; Fig. 2a,b) and was only borderline significant when all 115 replicate populations were included ($P = 0.072$).

Identifying contributors of divergent evolution between a pair of environments

The success of the multivariate Poisson in describing the distribution of mutation counts within a given environment and the overall weak signals of covariance provide the justification necessary to model the distribution of mutation counts among genes as an assemblage of effectively independent variables. We can then model divergent evolution at a given gene as the difference in 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 ($|\Delta n|$). The distribution of $|\Delta n|$ has been previously derived and is known as the Skellam distribution[18]. Starting with the null Poisson rates for each treatment ($\lambda_1 = n_{tot}^{(1)}/N_{genes}$; $\lambda_2 = n_{tot}^{(2)}/N_{genes}$), we define the probability mass function of the absolute value of $|\Delta n| = \left|n_i^{(1)} - n_i^{(2)}\right|$ as

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

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[19], we can define the P-value as

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

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

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

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 FDR α as

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

We then applied this approach to an experiment with replicate populations for four treatments[12]. We were able to identify genes that were consistently enriched for nonsynonymous mutations within a given treatment across all pairwise treatment comparisons (Table S1), largely agreeing with the conclusions of the original study[12].

Concluding Remarks

In this study, we investigated the distribution of mutation counts in evolve-and-resequence experiments. We found that a Poisson model sufficiently explains the distribution of mutation counts across genes. Using this result, we proposed that the difference in Poisson rates between treatments (i.e., the Skellam distribution) can be used to identify genes that contribute towards divergent evolution. These results can serve as a useful tool for analyzing the results of evolve-and-resequence experiments.

Methods

Predicting and quantifying parallelism

To determine the degree that we can predict statistical patterns of parallel evolution, we used a publicly available dataset 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.

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 in all empirical data by calculating the excess number of mutations (i.e., *multiplicity*) $m_{i,j} = n_{i,j} \cdot \frac{\bar{L}}{L_i}$, where L_i is the number of nonsynonymous sites in the ith gene and $\bar{L}$ is the mean of all genes in the genome[19]. To determine whether covariance can be reliably detected a given level of replication we estimated the largest normalized eigenvalue[20,21], defined as

$$\tilde{L}_1 = \frac{L_1 - \mu(n,g)}{\sigma(n,g)} \tag{7}$$

where L_1 is normalized as $L_1 = n\lambda_1 / \sum_{i=1}^{n} \lambda_i$ to sum to n and

$$\mu(n,g) = \frac{\left(\sqrt{g-1} + \sqrt{n}\right)^2}{g} \tag{8}$$

$$\sigma(n,g) = \frac{\sqrt{g-1} + \sqrt{n}}{g} \left(\frac{1}{\sqrt{g-1}} + \frac{1}{\sqrt{n}} \right)^{\frac{1}{3}} \tag{9}$$

As $n, g \to \infty$ and $n/g \to \gamma \geq 1$ $\tilde{L}_1$ tends towards a Tracy-Widom distribution[21,22]. Though these limits can be relaxed[20,23]. A null distribution of $\tilde{L}_1$ 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[24] of the ASA159 algorithm[25].

Available Code and Data

Instructions to obtain public data and code to reproduce our analyses are on GitHub: https://github.com/LennonLab/ParEvol.

Acknowledgments

This work was supported by US Army Research Office Grant W911NF-14-1-0411.

References

1. Gould, S. J. *Wonderful life: the Burgess Shale and the nature of history* (Norton & Co, New York, 1990).
2. Colosimo, P. F. *et al.* Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles. *Science* **307**, 1928–1933, DOI: 10.1126/science.1107239 (2005).
3. Bolnick, D. I., Barrett, R. D., Oke, K. B., Rennison, D. J. & Stuart, Y. E. (Non)Parallel Evolution. *Annu. Rev. Ecol. Evol. Syst.* **49**, 303–330, DOI: 10.1146/annurev-ecolsys-110617-062240 (2018).
4. Cooper, V. S. Experimental Evolution as a High-Throughput Screen for Genetic Adaptations. *mSphere* **3**, DOI: 10.1128/mSphere.00121-18 (2018).
5. McDonald, M. J. Microbial experimental evolution – a proving ground for evolutionary theory and a tool for discovery. *EMBO reports* **20**, e46992, DOI: 10.15252/embr.201846992 (2019). Publisher: John Wiley & Sons, Ltd.
6. Vogwill, T., Kojadinovic, M., Furió, V. & MacLean, R. C. Testing the Role of Genetic Background in Parallel Evolution Using the Comparative Experimental Evolution of Antibiotic Resistance. *Mol. Biol. Evol.* **31**, 3314–3323, DOI: 10.1093/molbev/msu262 (2014).
7. Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A. & Lenski, R. E. Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. *Proc. Natl. Acad. Sci.* **103**, 9107–9112, DOI: 10.1073/pnas.0602917103 (2006).
8. Bailey, S. F., Blanquart, F., Bataillon, T. & Kassen, R. What drives parallel evolution? *BioEssays* **39**, e201600176, DOI: 10.1002/bies.201600176 (2017).
9. Bertels, F., Leemann, C., Metzner, K. J. & Regoes, R. R. Parallel Evolution of HIV-1 in a Long-Term Experiment. *Mol. Biol. Evol.* **36**, 2400–2414, DOI: 10.1093/molbev/msz155 (2019). Publisher: Oxford Academic.
10. Fisher, K. J., Kryazhimskiy, S. & Lang, G. I. Detecting genetic interactions using parallel evolution in experimental populations. *Philos. Transactions Royal Soc. B: Biol. Sci.* **374**, 20180237, DOI: 10.1098/rstb.2018.0237 (2019).
11. Tenaillon, O. *et al.* The Molecular Diversity of Adaptive Convergence. *Science* **335**, 457–461, DOI: 10.1126/science.1212986 (2012).
12. Turner, C. B., Marshall, C. W. & Cooper, V. S. Parallel genetic adaptation across environments differing in mode of growth or resource availability. *Evol. Lett.* **2**, 355–367, DOI: 10.1002/evl3.75 (2018).
13. Shoemaker, W. R. *et al.* Microbial population dynamics and evolutionary outcomes under extreme energy-limitation. *bioRxiv* DOI: 10.1101/2021.01.25.428163 (2021). Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2021/01/26/2021.01.25.428163.full.pdf.
14. Shoemaker, W. R., Polezhaeva, E., Givens, K. B. & Lennon, J. T. Molecular evolutionary dynamics of energy limited microorganisms. *bioRxiv* DOI: 10.1101/2021.02.08.430186 (2021). https://www.biorxiv.org/content/early/2021/02/24/2021.02.08.430186.full.pdf.
15. Tenaillon, O. *et al.* Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* **536**, 165–170, DOI: 10.1038/nature18959 (2016).
16. Good, B. H. & Hallatschek, O. Effective models and the search for quantitative principles in microbial evolution. *Curr. Opin. Microbiol.* **45**, 203–212, DOI: 10.1016/j.mib.2018.11.005 (2018).
17. Grilli, J. Laws of diversity and variation in microbial communities. *bioRxiv* 680454, DOI: 10.1101/680454 (2019). Publisher: Cold Spring Harbor Laboratory Section: New Results.
18. Skellam, J. G. The frequency distribution of the difference between two poisson variates belonging to different populations. *J. Royal Stat. Soc. Ser. A (General)* **109**, 296 (1946).
19. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50, DOI: 10.1038/nature24287 (2017).
20. Tracy, C. A. & Widom, H. Level-spacing distributions and the airy kernel. *Comm. Math. Phys.* **159**, 151–174 (1994).
21. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLOS Genet.* **2**, e190, DOI: 10.1371/journal.pgen.0020190 (2006).
22. Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **29**, 295–327, DOI: 10.1214/aos/1009210544 (2001).

23. Soshnikov, A. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J. Stat. Phys.* **108**, 1033–1056, DOI: 10.1023/A:1019739414239 (2002).

24. Baak, M., Koopman, R., Snoek, H. & Klous, S. A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics (2019). 1811.11440.

25. Patefield, W. M. Algorithm as 159: An efficient method of generating random r × c tables with given row and column totals. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **30**, 91–97, DOI: 10.2307/2346669 (1981). Publisher: [Wiley, Royal Statistical Society].

Figures

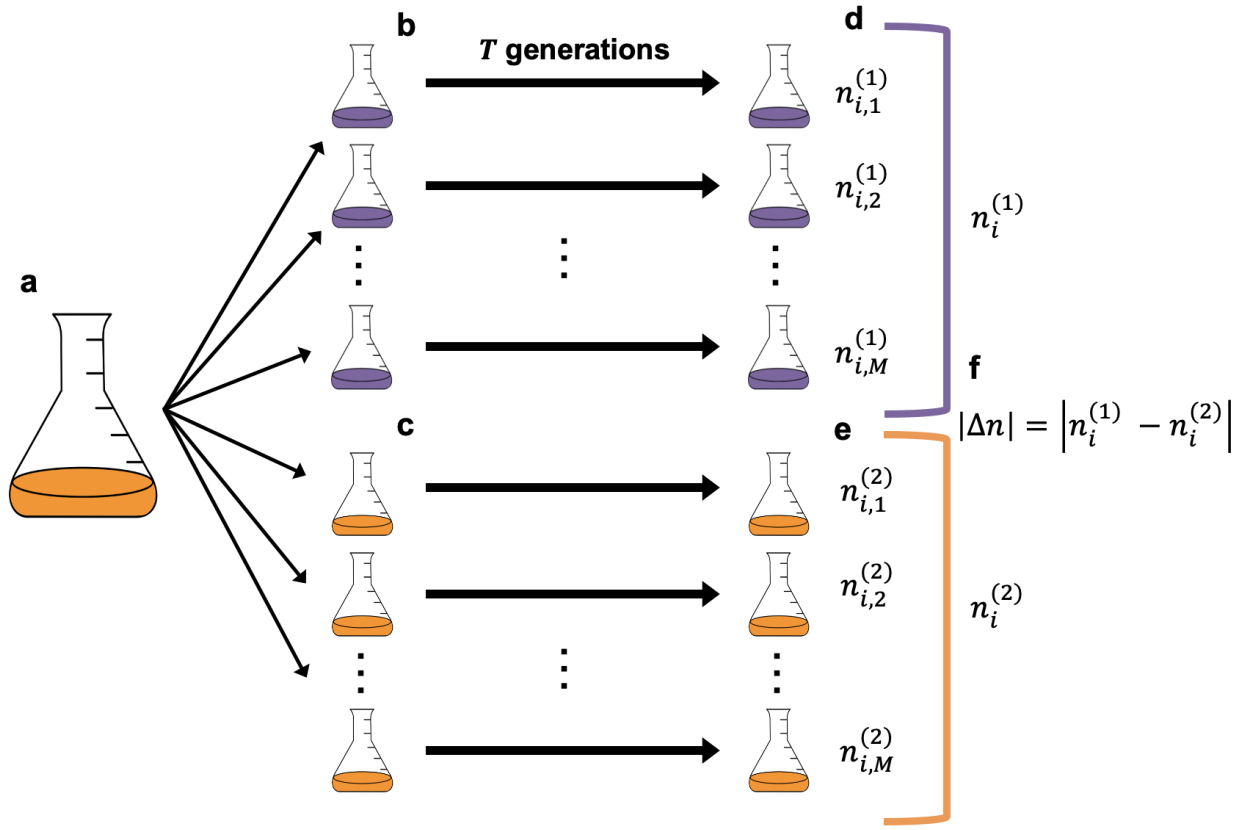

Figure 1. a) A typical evolve-and-resequence experiment is performed by splitting a culture grown from a single colony inoculate into replicate flasks constituting one or more environment (e.g., purple or orange) and propagating the culture over time. **b,c)** After a given number of generations has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

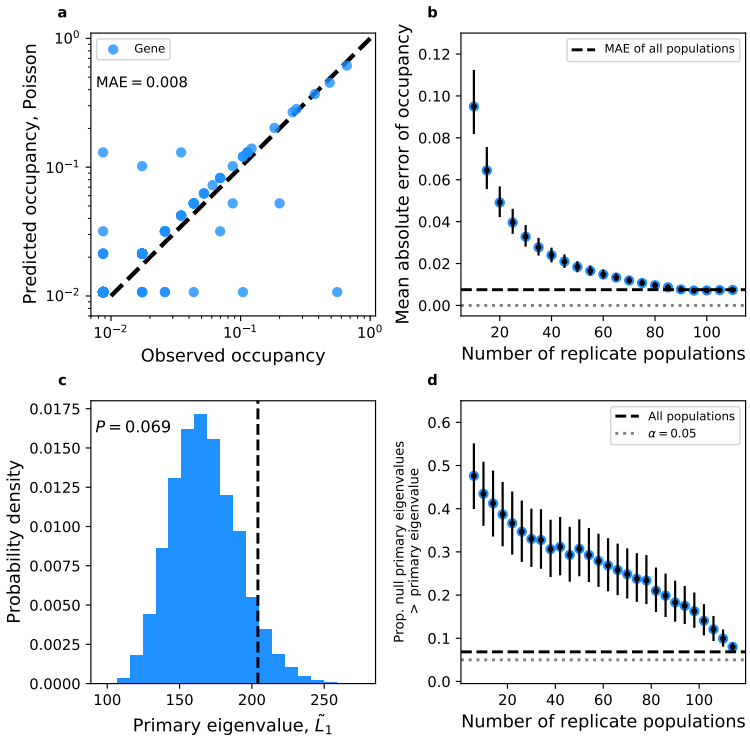

Figure 2. **a)** Using the Poisson distribution, we can predict the occupancy of nonsynonymous mutations for a given gene among 115 replicate *E. coli* populations. **b)** The amount of error rapidly decreases as the number of replicate populations increases. **c)** The degree of covariance in a gene-by-population matrix can be summarized by the primary eigenvalue (dashed black line). By generating null count matrices, we can calculate a null distribution of primary eigenvalues and calculate a *P*-value. **d)** By subsampling replicate populations without replacement, we can calculate the fraction of observed primary eigenvalues greater than the null.

Supplemental material

Treatment	Locus tag	Function
High carbon, bead	BCEN2424_RS08045	Type 1 fimbrial protein
High carbon, planktonic	BCEN2424_RS08125	Lysine N(6)-hydroxylase
Low carbon, bead	BCEN2424_RS16880	MFS transporter
	BCEN2424_RS25830	FkbM family methyltransferase
Low carbon, planktonic	BCEN2424_RS03065	IclR family transcriptional regulator
	BCEN2424_RS04940	YicC family protein

Table S1. Using Eq. 4, we calculated a *P*-value of divergent evolution for each pairwise treatment comparison for each gene. To identify candidates of adaptation that are unique to a given treatment, we identified the set of genes that were significantly enriched for nonsynonymous mutations within a given treatment for all pairwise treatment comparisons.