Quantifying Parallel Evolution

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

[1]Department of Biology, Indiana University, Bloomington, IN, 47405, USA

[*]Correspondence to: wrshoema@indiana.edu

May 13, 2020

Abstract

Parallel evolution is consistently observed across the tree of life. However, the degree of parallelism between replicate populations in evolution experiments is rarely quantified at the gene level. Here we examine parallel evolution as the degree of covariance between replicate populations, providing a justification for the use of dimensionality reduction. We examine the extent that signals of gene-level covariance can be inferred in microbial evolve-and-resequence evolution experiments, finding that deviations from parallelism are difficult to quantify at a given point in time. However, this low statistical signal means that covariance between replicate populations is unlikely to interfere with the ability to detect divergent evolutionary trajectories for populations in different environments. Finally, we find evidence suggesting that temporal patterns of parallelism are comparatively easier to detect and that these patterns may reflect the evolutionary dynamics of microbial populations.

Keywords— Experimental evolution, Microbial evolution, Parallel evolution

1 Introduction

Parallel evolution occurs when independent populations evolve similar phenotypes and genotypes. Observed across the tree of life [12, 41, 27], parallel evolution has historically been viewed as a singular outcome that is representative of adaptation [24]. However, parallelism is not binary [8, 47, 32]. Instead, parallelism is a continuous quantity that captures the variation in evolutionary outcomes, allowing for researchers to test hypotheses about the extent that evolutionary and ecological forces affect the repeatability of evolutionary outcomes relative to a null expectation.

The idea that parallelism should be viewed as a quantity is particularly suited to the experimental study of microbial evolution, where many large populations with short generation times can be simultaneously maintained. In microbial systems the same evolutionary outcome can repeatedly occur across levels of biological organization, ranging from nucleotide sites repeatedly acquiring the same mutation [7] to phenotypes consistently changing in the same direction and magnitude [17] to predator-prey systems repeatedly evolving similar dynamics [18]. The experimental tractability of many microbial systems also allows for the degree of parallelism to be examined across diverse ecological scenarios. For example, it has been argued that an exceptional degree of parallel outcomes has been observed in evolution experiments where microbial populations adapt to high temperatures [50], alternative resources [21], and the introduction of new species [45]. The power of experimental microbial evolution provides unique opportunities for the degree of variation in evolutionary outcomes to be examined across biological hierarchies and environments.

Parallel evolution can be found across biological scales, though it is not equally likely at each scale. Independently evolving bacterial populations are unlikely to acquire mutations at the same nucleotide site in most evolve-and-resequence experiments [13], making it necessary to group mutations together. Under this coarse-graining, genetic parallelism is examined as the set of genes that acquire more mutations than expected by chance. A number of statistical approaches have been developed and applied to evolution experiments to identify this set of genes [55, 6, 49, 23, 3]. In

addition, in recent years increases attention has been given to the shape of this distribution of mutations across genes, with a particular focus on developing a reasonable statistical null for parallelism [23] and identifying evolutionary mechanisms that drive the shape of the distribution [4].

While the distribution of mutations among genes has been given considerable attention, relatively few attempts have been made to examine the joint distribution of mutation counts between genes [15]. Epistatic interactions between mutations in different genes make certain combinations of mutation counts more likely than others, generating covariance between populations [5] (analogous to within/between population genetic variation [20] or α/β species diversity [56]). Conceptually, this covariance can be understood as the inverse of parallel evolution, where higher levels of covariance between genes makes replicate populations less genetically similar. Because more genes acquire mutations than there are replicate populations for the vast majority of evolution experiments, dimensionality reduction is often necessary to determine whether covariance exists. Dimensionality reduction approaches have been applied to determine whether replicate populations in different environments diverged at the gene level [52], though these approaches have yet to be used to quantify the degree of parallelism among replicate populations .

Here, we examine how covariance between genes relates to the experimental evolution of microbial populations. We investigate how a stochastic formulation of Principal Component Analysis [44] relates to covariance between genes and how that covariance can be accounted for to determine whether the outcome of an evolution experiment was more or less parallel than expected by chance. We argue that in the context of experimental evolution the concept of parallelism should be treated as a continuous quantity where the absence of covariance between genes represents a statistical null to be rejected. We compare mathematical approaches from statistical physics and multivariate statistics using simulations to quantify the degree of parallelism and its significance. We then examine whether deviations from parallelism interfere with the ability to detect divergent evolution in case studies where replicate populations evolved under different conditions. Finally, we examine how parallelism varies over time in a

highly temporally resolved evolution experiment.

2 Materials and Methods

2.1 Parallel evolution and PCA

We examine the relationship between conceptualizations of parallel evolution and PCA. We assume that n replicate populations have been propagated for an equal number of generations in the same environment. Assuming that the populations are evolving under the strong selection, weak mutation limit (SSWM), the molecular dynamics can be examined as a biased random walk on genotypic space consisting of L biallelic sites that comprises the set of epistatic interactions between sites. Once populations have been sequenced, a site-by-population matrix can be constructed, where each value represents the presence or absence of a given mutation in a given population. While there is evidence that parallel outcomes can occur at the nucleotide level in microbial evolution experiments [23], it is far more common in organisms with smaller genomes and larger population sizes such as viruses [7]. Instead, to examine parallelism, it is reasonable to reduce sparsity by constructing an $G \times n$ population-by-gene count matrix $\mathbf{Z}$, effectively coarse-graining genotypic space into G genes. At this point the question of whether or not parallelism is present in an evolution experiment can be understood as the degree that epistatic interactions between sites translates to an observable statistical signal at the gene-level.

To understand how $\mathbf{Z}$ relates to the concept of parallelism is it useful to use PCA as a conceptual intermediate. If elements of $\mathbf{Z}$ have been centered by the mean of each column as $X_{i,j} = Z_{i,j} - \frac{1}{n}\sum_{k=1}^{n} Z_{i,k}$ to create the zero-centered matrix $\mathbf{X}$, then the empirical population covariance matrix can be estimated as

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T \tag{1}$$

The principal components of $\mathbf{X}$ are obtained from the eigenvectors of $\mathbf{C}$. However, PCA is closely connected to the factorization process of Singular Value Decomposition

(SVD) [44], which has been previously used to establish intuitive connections between evolutionary processes and PCA [34]. Following this approach, the SVD is performed using the stochastic matrix $\mathbf{M}$:

$$\mathbf{M} = \frac{1}{G}\mathbf{X}^T\mathbf{X} \tag{2}$$

As $\mathbf{M}$ is a stochastic matrix, the expected value for each element can be examined as:

$$\mathrm{E}\left[M_{i,j}\right] = \frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[X_{g,i}X_{g,j}\right] \tag{3a}$$

$$= \frac{1}{G}\sum_{g}^{G}\mathrm{E}\left[\left(Z_{g,i} - \frac{1}{n}\sum_{k=1}^{n}Z_{g,k}\right)\left(Z_{g,j} - \frac{1}{n}\sum_{k=1}^{n}Z_{g,k}\right)\right] \tag{3b}$$

By expanding the brackets, the expected value of $M_{i,j}$ for a single gene g is

$$\mathrm{E}\left[M_{i,j}^{(g)}\right] = \mathrm{E}\left[Z_{g,i}Z_{g,j}\right] - \frac{1}{n}\sum_{k=1}^{n}\mathrm{E}\left[Z_{g,i}Z_{g,k}\right] - \frac{1}{n}\sum_{k=1}^{n}\mathrm{E}\left[Z_{g,j}Z_{g,k}\right] + \sum_{k=1}^{n}\sum_{l=1}^{n}\mathrm{E}\left[Z_{g,k}Z_{g,l}\right] \tag{4}$$

Each element of eqn. 4 contains at least one expected value of two joint random variables, which can be viewed as the sum of the products of the expected value of each random variable and their covariance (ex., $\mathrm{E}\left[Z_{g,i}Z_{g,j}\right] = \mathrm{E}\left[Z_{g,i}\right]\mathrm{E}\left[Z_{g,j}\right] + \mathrm{cov}\left(Z_{g,i}Z_{g,j}\right)$). Assuming that no cross-contamination occurred over the course of the experiment, our populations are evolutionarily independent and we can set $\mathrm{cov}\left(Z_{g,i}Z_{g,j}\right) = 0$.

We note that this covariance term can in principle be modified to account for shared evolutionary history in experimental evolutionary studies where multiple taxa with a resolved phylogeny have evolved in the same environment. More importantly, because our populations are independent, under a SSWM limit the presence of between-population covariance values greater than expected by chance indicates the presence of epistatic interactions. Therefore, the concept of absolute parallelism between populations in experimental evolution relates to PCA as the absence of covariance between genes, a null expectation that can be statistically tested.

2.2 Signals of non-parallelism

Random versions of $\mathbf{Z}$ ($\mathbf{Z}^*$) were obtained by randomizing the co-occurrence of mutations across genes. We chose to generate $\mathbf{Z}^*$ such that row and column sums are conserved, an approach that reduces covariance between genes while conserving the observed distribution of evolutionary distances and the distribution of per-gene mutation counts, respectively. This was done by adapting previously developed Python code [37] and the `ASA159 FORTRAN77` library [38].

Deviations from parallelism were quantified using statistics frequently used in random matrix theory and multivariate statistical testing. The first two statistics are commonly used for analyses in ordination space, specifically the principal components (PCs) for the purpose of this study. The first statistic is the Mean Centroid Distance (MCD), a common measure of dispersion defined as

$$\text{MCD} = \frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{k}\left|P_{i,j} - \overline{p}_j\right|^2\right)^{\frac{1}{2}} \tag{5}$$

where $\mathbf{P}^{(k)}$ is the $n \times k$ matrix consisting of the first k principal axes and $\overline{p}_j$ is the mean of the jth axis [30].

The second statistic is the Mean Pairwise Distance (MPD), a statistic frequently used when comparing variation within and between groups in ordination space [2]. MPD is defined as

$$\text{MPD} = \frac{2}{n(n-1)}\sum_{i=2}^{n}\sum_{i=1}^{i-1} d(\mathbf{p}_i^{(k)}, \mathbf{p}_j^k) \tag{6}$$

where $\mathbf{p}_i^{(k)}$ is the k-element vector of the ith population and $d()$ is the Euclidean distance

The final statistic is the largest normalized eigenvalue [51, 39], 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 [29, 39]. Though these criteria can be relaxed [51] and $\tilde{L}_1$ holds for matrices as small as 5×20. This approach was initially developed for Wishart matrices with Gaussian distributed entries. While mutation counts in $\mathbf{X}$ are likely non-Gaussian, this is not critical and our data are unlikely to violate previously established criteria [46]. While this statistic is less frequently used in multivariate ecological and evolutionary analyses, we chose to include it due to the fact that the distribution of primary eigenvalues has analytic forms for certain classes of square matrices and is an active area of mathematical research [48], providing added interpretability to the statistic.

2.3 Quantifying parallelism in simulated data

While little is known about the distribution of gene-specific substitution rates, we are primarily interested in the covariance between genes that ultimately generates covariance between populations, so that the choice of a distribution that reflects the mean rate of evolution is not necessarily pertinent to examine the covariance. Therefore, we chose to generate the vector $\mathbf{g}$ containing G gene-specific substitution rates using a gamma distribution with a shape parameter of 3 and a scale parameter of 1. To generate the between-gene covariance matrix we first generated scale-free random graphs using the Barabási-Albert preferential attachment model [1]. The `barabasi_albert_graph` and the `powerlaw_cluster_graph` functions from the `networkx` Python package [36] were used to generate Barabási-Albert graphs and clustered Barabási-Albert graphs [26], respectively. The adjacency matrix of the graph was multiplied by a given covariance value and the diagonal elements were set to one so that the matrix fit the standard normal form ($\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$). We only proceeded

with the simulation if Σ was positive definite, the probability of which decreases with increasing values of σ under the Geršgorin circle theorem [43]. Poisson distributed mutation counts were generated using inverse transform sampling [14] with the cutoff determined by samples of the Cumulative Density Function of $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ rather than the standard approach of sampling from a uniform distribution $\mathcal{U}(\mathbf{0}, \mathbf{1})$ so that between gene covariance could be conserved (extended description in Supporting Information). PCA was performed using the `decomposition.PCA()` function from `scikit-learn` [40] in Python 3.6. Values from simulated $\mathbf{Z}$ matrices were compared to a null distribution of values calculated from 1,000 iterations of $\mathbf{Z}^*$. This process was repeated 1,000 times to estimate statistical power as the proportion of simulations where the null could be rejected at a significance level of $\alpha = 0.05$.

2.4 Quantifying parallelism in empirical data

To determine the degree that deviations from parallelism can be detected we used a publicly available data set from one of the largest microbial evolution experiments. In this experiment, 115 replicate populations of *Escherichia coli* were serially transferred for 2,000 generations at 42.2 °C [50]. A single colony was isolated from each replicate population and sequenced. We merged all mutations from all replicate populations into a single 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*) $m_{g,i} = Z_{g,i} \cdot \frac{\bar{L}}{L_g}$, where $\bar{L}$ is the mean size of all genes in the genome [23]. To measure the degree that reducing covariance affected clustering we calculated the variance ratio criteria using the Calinski and Harabaz score [11] on k-means clustered PC space [25] using `scikit-learn` [40]. Cluster stability was assessed by re-sampling populations in PC space with replacement, performing spectral clustering [25], and mapping clusters between the original and re-sampled PC space by their maximum Jaccard coefficient [33]. This process was repeated 10,000 times.

We compared our PCA-based results using data from [50] to analyses that do not account for covariance between genes. To do this, we summed across the rows of the population-by-gene matrix to generate a vector of the total number of mutations

acquired in each gene (n_i) and calculated multiplicity of each of the i genes as $m_i = n_i \cdot \frac{\bar{L}}{L_i}$. Values of m_i were compared to the null expectation of $\overline{m} = n_{tot}/N_{genes}$, where N_{genes} is the total number of genes in the genome, as the net increase in log-likelihood

$$\Delta\ell = \sum_i = n_i \log\left(\frac{m_i}{\overline{m}}\right) \tag{10}$$

Where probability values that a given gene has an excess number of mutations with a False Discovery Rate (FDR) of 0.05 were calculated for each gene as previously described [23]. We calculated the $\Delta\ell$, the number of significant genes, and the proportion of times that genes of interest had a significant multiplicity by sampling a given number of populations without replacement 10,000 times.

To examine the degree that covariance between replicates affects the ability to distinguish between populations evolving under different conditions, we examined two datasets from studies with moderate within-treatment replication. The first dataset examined the spectrum of mutations in genomically recoded *E. coli* MG1655, where 14 replicate populations of the following strains were serially transferred: (1) the non-recoded ancestor (ECNR2), (2) a strain where UAG stop codons were replaced with UAA and the class I peptide release factor 1 was deleted (C321.ΔA), (3) a C321.ΔA derivative with engineered reversions to three off-target mutations (C321.ΔA-v2), and (4) a C321.ΔA derivative recoded to restore RF1 (C321) [53]. The second study was more focused on the consequence of microbial life cycles in different environments. In this experiment *Burkholderia cenocepacia* with planktonic or biofilm life in environments containing with low or high concentrations of carbon [52]. The degree of evolutionary divergence was quantified using two forms of Permutational ANOVA (PERMANOVA) F statistics, a standard one (F_1) and one that accounts for unequal levels of parallelism among treatments (F_2) [2]. Null population-by-gene count matrices for each study were constructed for k treatments, randomized, and concatenated as $\mathbf{Z}^* = (\mathbf{Z}_1^*, \cdots, \mathbf{Z}_k^*)^{\mathrm{T}}$. All entries were relativized by dividing each element by the sum of its row.

To examine temporal trends in covariance between populations we used publicly available sequence data from the Long-term Evolution Experiment [31], an experiment

consisting of twelve *E. coli* populations that have been serially propagated for over 60,000 generations. We generated a population-by-gene count matrix every 500 generations for fixed mutations inferred in [23] and concatenated observations as a single matrix. We chose to only examine the six nonmutator populations: Ara+1, +2, +4, +5, -5 and -6, as hypermutator populations exhibit qualitatively different molecular dynamics [23] that could affect the covariance between populations. While there are a variety of geometric techniques to examine temporal patterns in ordination space [10], we elected the straightforward approach of randomizing timepoints for each replicate population so that null values of MPD could be estimate in the absence of temporal autocorrelation. The same multiplicity calculation was performed as described above. While there are a number of techniques to estimate the number of PCs to keep [42, 9, 19, 35], we elected to keep a number of PCs equal to the number of replicate populations for the LTEE data.

3 Results

3.1 Gene-level covariance is low

We find that statistical power for rejecting the null hypothesis of zero covariance between genes ($H_0 : \mathbf{\Sigma} = \mathbf{I}$) increases with covariance, but is generally low with the probability only reaching 0.25 with the highest covariance examined (Fig. 1). The statistics MCD and MPD calculated on the first principal component have much lower power than the more commonly used statistic $\tilde{L}_1$, though they overtake $\tilde{L}_1$ once additional PCs are considered. Given that the statistics were fairly similar and that MPD is used to calculate F_2 [2], we used MPD for the remaining analyses. Statistical power slightly increases with the degree of clustering, though the increment is very small for the range of clustering coefficients examined (Fig. 1) which indicates that the structure of the between-gene covariance matrix does not influence our ability to detect covariance between populations. Similar patterns were observed for the effect size (standardized score; Fig. 1). Though the ability to reject the null hypothesis requires a large number of replicate populations as well as a large number of genes

that acquire mutations (Fig. S1)

We find clear evidence of population covariance in existing data [50]. The *E. coli* populations appear to form three clusters in PC space (Fig. 2), where the formation of the two smaller clusters are primarily driven by mutations acquired in ESCRE1901 and ECB_01992 along the first and second principle components, receptively. Both genes are putative proteins with no known function that have acquired mutations in separate evolution experiments examining *E. coli* adaptation to heat [28]. Of all genes, ESCRE1901 has the highest squared correlation with the first principal component (i.e., rescaled loading; $\rho^2 = 0.92$), the same being true for ECB_01992 and the second PC ($\rho^2 = 0.73$).

We find that the observed $\text{MPD}^{(3)}$ is significantly greater than the null expectation in the absence of covariance (Fig. 2, S2), though, consistent with our simulations (Fig. S1), the required replication to consistently reject the null is over an order of magnitude larger than the replication level of most standard evolution experiments (Fig. 2). This pattern holds at the gene level, as similar replication is needed to determine if ESCRE1901 and ECB_01992 acquire more mutations than expected by chance across all replicate populations (3). That cluster formation is driven by a few genes explains the low stability of the clusters (Fig. 2), despite the fact that the variance ratio between and within clusters is much higher than what is found in null count matrices (Fig. 2). That few genes (and, therefore, few mutations) drive this covariance explains the lack of a clear relationship between either of the first two PCs or clusters in PC space and the relative fitness of each clone (Fig. S3).

3.2 Within-group covariance does not interfere with the ability to detect divergence.

We find no significant difference between observed MPD values and the null expectation when covariance is removed from the population-by-gene matrix of each treatment in two evolution experiments with multiple treatments and moderate replication (Fig. 4, S4). This pattern holds at the level of summary statistics, as there is no significant

difference between estimates of between vs. within treatment variation and the null expectation in the absence of covariance for either F statistic (Fig. 4, Fig. S5).

3.3 Temporal patterns of parallelism are detectable at the gene-level

Our previous results suggest that it would be difficult to infer whether there was a significant amount of between-gene covariance at a given timepoint in evolution experiments with a standard number of replicate populations. Indeed, that is also the case for the LTEE (Fig. S6). Instead, we chose to examine how MPD varied over time. In contrast with our attempts to detect covariance at a single time point, there are clear temporal patterns of parallelism in the LTEE despite there only being six replicate populations. While it is trivial that the genetic distance between initially identical replicate populations grown from a single clone has to increase, we see that after a period of increasing distance the replicate populations begin to become more similar (Fig. 5). By measuring MPD over the first five axes ($\text{MPD}^{(5)}$, Fig. S7), we find that there is a clear pattern where $\text{MPD}^{(5)}$ rapidly increases over the first few thousand generations and gradually decreases starting at 4,750 generations.

4 Discussion

Our results suggest that it is difficult to detect covariance between populations at the gene-level in evolve-and-resequence evolution experiments with a standard level of replication. A minimum of 60 replicate populations are required to reject the null hypothesis of zero covariance 50% of the time in [50]. This may in part be due to the fact that individual clones were sequenced in this experiment, whereas pooled sequencing would provide estimates of mutation frequencies which may contain additional information about their fitness effects. However, the number of replicate populations required was similar to our results from simulated data, suggesting that covariance cannot be detected at the gene level in the vast majority of evolution experiments.

While covariance was weak, we were able to identify genes that disproportionately

contribute to the observed signal. Covariance between populations in [50] is primarily driven by ESCRE1901 and ECB_01992, two genes of unknown function that have also acquired mutations in a similarly designed experiment [28]. Given that covariance can indicate the presence of an interaction, ESCRE1901 and ECB_01992 are useful candidates for investigating between-gene epistatic interactions in *E. coli*. However, there is no relationship between fitness and gene-level mutational composition or the presence of mutations in these genes. This lack of a relationship may be the result of the mutations in these genes making a relatively small overall contribution to fitness that cannot be detected at a coarse scale, as suggested by the fact that 50 replicate populations are required to determine that ESCRE1901 and ECB_01992 acquire more mutations than expected by chance 95% of the time.

Observed F statistics were not significantly different from the null expectation in absence of within-group covariance for the datasets examined [52, 53]. This result suggests that while covariance between populations is difficult to detect in evolution experiments with moderate replication (e.g., n=4-6), this low signal provides the added advantage of not having to be concerned with how different environments or backgrounds affect covariance between genes (i.e., the Behrens–Fisher problem [16, 54]). Rather, the difference in mean gene-level substitution rates between treatments is likely greater than the covariance. While the experiments we examined were conducted in disparate environments or with synthetic strains, we argue that these conclusions will hold for experiments that examine microbial evolution across a more continuous environmental or genetic gradient.

While covariance between populations does not interfere with the ability to detect divergent evolution, we find evidence that covariance between replicate populations changes over time. In the LTEE we find that MPD rapidly increases over the first 4,750 generations, followed by a steady decrease over the remaining 55,000 generations. This pattern is consistent with the "two-epoch" mean-field model of adaptation that has been proposed for this system, where populations evolve under an initial burst of macroscopic epistasis followed by the steady accumulation of mutations under a constant distribution of fitness effects [22]. That is, qualitative shifts in underlying

evolutionary dynamics may be detectable by examining covariance at the gene-level over time. While this transition between regimes has been suggested to occur at the 10,000 generation mark [22], the difference of a few thousand generations does not negate the presence of the qualitative trend and this result may be corroborated by examining how gene-level interactions give rise to evolutionary dynamics predicted by mean-field models.

As long-term experiments become increasingly used to examine evolutionary dynamics and test hypotheses it is necessary to identify appropriate statistical approaches and establish their limitations. Our work suggests that ordination techniques have a number of potential applications for experimental evolution. PCA specifically has the added advantage of being a well understood statistical tool for examining covariance, which can be connected to the joint probability distribution of gene-level substitution rates. The structure of the covariance between genes is ultimately of theoretical interest and while our results suggest that its statistical signal is small and the population-by-gene matrix is sparse, we are able to identify contributing genes with sufficient replication and identify temporal trends. For the more complex case of covariance over time, it will be necessary to examine this joint distribution in greater detail by incorporating it into models of evolutionary dynamics.

5 Author Contributions

WRS and JTL conceived the experiments and wrote the paper. WRS designed and performed the experiments and analysed the data.

6 Acknowledgements

We thank Benjamin H. Good and Michael M. Desai for their insightful comments at an early stage of this work. Financial support was provided by the National Science Foundation (1442246, JTL), US Army Research Office (W911NF-14-1-0411, JTL), and National Aeronautics and Space Administration (80NSSC20K0618, JTL). Computing

resources for simulations was supported by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, the National Science Foundation under Grant No. CNS-0521433, and Shared University Research grants from IBM, Inc., to Indiana University.

7 Data Archiving

No new empirical data was generated for this study. Reproducible code to perform the analyses in this study is available on GitHub as: https://github.com/LennonLab/ParEvol. Simulated data is available on Zenodo as DOI: 10.5281/zenodo.3779341.

References

[1] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Rev. Mod. Phys.* 74 (1 Jan. 2002), pp. 47–97. DOI: 10.1103/RevModPhys.74.47. URL: https://link.aps.org/doi/10.1103/RevModPhys.74.47.

[2] Marti J. Anderson et al. "Some solutions to the multivariate Behrens–Fisher problem for dissimilarity-based analyses". en. In: *Australian & New Zealand Journal of Statistics* 59.1 (2017), pp. 57–79. ISSN: 1467-842X. DOI: 10.1111/anzs.12176. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/anzs.12176 (visited on 01/15/2020).

[3] Susan F. Bailey, Qianyun Guo, and Thomas Bataillon. "Identifying Drivers of Parallel Evolution: A Regression Model Approach". en. In: *Genome Biology and Evolution* 10.10 (Oct. 2018), pp. 2801–2812. DOI: 10.1093/gbe/evy210. URL: https://academic.oup.com/gbe/article/10/10/2801/5106663 (visited on 01/21/2020).

[4] Susan F. Bailey et al. "What drives parallel evolution?" en. In: *BioEssays* 39.1 (2017), e201600176. ISSN: 1521-1878. DOI: 10.1002/bies.201600176. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201600176 (visited on 01/21/2020).

[5] Benedikt Bauer and Chaitanya S. Gokhale. "Repeatability of evolution on epistatic landscapes". en. In: *Scientific Reports* 5.1 (May 2015), pp. 1–6. ISSN: 2045-2322. DOI: 10.1038/srep09607. URL: https://www.nature.com/articles/srep09607 (visited on 01/21/2020).

[6] Megan G. Behringer et al. "Escherichia coli cultures maintain stable subpopulation structure during long-term evolution". en. In: *Proceedings of the National Academy of Sciences* 115.20 (May 2018), E4642–E4650. ISSN:

0027-8424, 1091-6490. DOI: 10.1073/pnas.1708371115. URL: https://www.pnas.org/content/115/20/E4642 (visited on 01/20/2020).

[7] Frederic Bertels et al. "Parallel Evolution of HIV-1 in a Long-Term Experiment". In: *Molecular Biology and Evolution* 36.11 (2019), pp. 2400–2414. ISSN: 0737-4038. DOI: 10.1093/molbev/msz155. URL: https://doi.org/10.1093/molbev/msz155.

[8] Daniel I. Bolnick et al. "(Non)Parallel Evolution". In: *Annual Review of Ecology, Evolution, and Systematics* 49.1 (2018), pp. 303–330. DOI: 10.1146/annurev-ecolsys-110617-062240. URL: https://doi.org/10.1146/annurev-ecolsys-110617-062240 (visited on 01/20/2020).

[9] R. Bro et al. "Cross-validation of component models: a critical look at current methods". eng. In: *Analytical and Bioanalytical Chemistry* 390.5 (Mar. 2008), pp. 1241–1251. ISSN: 1618-2650. DOI: 10.1007/s00216-007-1790-1.

[10] Miquel De CÃ¡ceres et al. "Trajectory analysis in community ecology". en. In: *Ecological Monographs* 89.2 (2019), e01350. ISSN: 1557-7015. DOI: 10.1002/ecm.1350. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1350 (visited on 01/20/2020).

[11] T. Calinski and J. Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics* 3.1 (Jan. 1974), pp. 1–27. ISSN: 0090-3272. DOI: 10.1080/03610927408827101. URL: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101 (visited on 01/25/2020).

[12] Pamela F. Colosimo et al. "Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles". In: *Science* 307.5717 (2005), pp. 1928–1933. ISSN: 0036-8075. DOI: 10.1126/science.1107239. eprint: https://science.sciencemag.org/content/307/5717/1928.

425 full.pdf. URL: https://science.sciencemag.org/content/307/
426 5717/1928.

427 [13] Vaughn S. Cooper. "Experimental Evolution as a High-Throughput Screen
428 for Genetic Adaptations". In: *mSphere* 3.3 (2018). Ed. by Ana Cristina
429 Gales. DOI: 10.1128/mSphere.00121-18. URL: https://msphere.asm.
430 org/content/3/3/e00121-18.

431 [14] Luc Devroye. "Discrete Univariate Distributions". en. In: *Non-Uniform
432 Random Variate Generation*. New York, NY: Springer New York, 1986,
433 pp. 485–553. ISBN: 978-1-4613-8645-2. DOI: 10.1007/978-1-4613-8643-
434 8_10. URL: http://link.springer.com/10.1007/978-1-4613-8643-
435 8_10 (visited on 01/12/2020).

436 [15] Kaitlin J. Fisher, Sergey Kryazhimskiy, and Gregory I. Lang. "Detecting
437 genetic interactions using parallel evolution in experimental populations".
438 In: *Philosophical Transactions of the Royal Society B: Biological Sciences*
439 374.1777 (July 2019), p. 20180237. DOI: 10.1098/rstb.2018.0237. URL:
440 https://royalsocietypublishing.org/doi/full/10.1098/rstb.
441 2018.0237 (visited on 01/26/2020).

442 [16] R. A. Fisher. "The Fiducial Argument in Statistical Inference". en. In:
443 *Annals of Eugenics* 6.4 (1935), pp. 391–398. ISSN: 2050-1439. DOI: 10.
444 1111/j.1469-1809.1935.tb02120.x. URL: https://onlinelibrary.
445 wiley.com/doi/abs/10.1111/j.1469-1809.1935.tb02120.x (visited
446 on 01/28/2020).

447 [17] David T Fraebel et al. "Environment determines evolutionary trajectory in
448 a constrained phenotypic space". In: *eLife* 6 (Mar. 2017). Ed. by Wenying
449 Shou, e24669. ISSN: 2050-084X. DOI: 10.7554/eLife.24669. URL: https:
450 //doi.org/10.7554/eLife.24669 (visited on 01/20/2020).

[18] Jens Frickel et al. "Population size changes and selection drive patterns of parallel evolution in a host-virus system". en. In: *Nature Communications* 9.1 (Apr. 2018), pp. 1–10. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03990-7. URL: https://www.nature.com/articles/s41467-018-03990-7 (visited on 02/09/2020).

[19] Matan Gavish and David L. Donoho. "The Optimal Hard Threshold for Singular Values is $4/\sqrt 3$". In: *IEEE Transactions on Information Theory* 60.8 (Aug. 2014), pp. 5040–5053. ISSN: 1557-9654. DOI: 10.1109/TIT.2014.2323359.

[20] John H. Gillespie. *Population Genetics: A Concise Guide*. English. 2nd edition. Baltimore, Md: JHUP, Aug. 2004. ISBN: 978-0-8018-8009-4.

[21] Shmuel Gleizer et al. "Conversion of Escherichia coli to Generate All Biomass Carbon from CO2". en. In: *Cell* 179.6 (Nov. 2019), 1255–1263.e12. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.11.009. URL: http://www.sciencedirect.com/science/article/pii/S0092867419312309 (visited on 01/20/2020).

[22] Benjamin H. Good and Michael M. Desai. "The Impact of Macroscopic Epistasis on Long-Term Evolutionary Dynamics". en. In: *Genetics* 199.1 (Jan. 2015), pp. 177–190. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.114.172460. URL: https://www.genetics.org/content/199/1/177 (visited on 01/26/2020).

[23] Benjamin H. Good et al. "The dynamics of molecular evolution over 60,000 generations". en. In: *Nature* 551.7678 (Nov. 2017), pp. 45–50. ISSN: 1476-4687. DOI: 10.1038/nature24287. URL: https://www.nature.com/articles/nature24287 (visited on 01/21/2020).

[24] Paul H. Harvey and Mark D. Pagel. *The comparative method in evolutionary biology*. Oxford series in ecology and evolution. Oxford ; New York: Oxford University Press, 1991. ISBN: 0198546408.

[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. en. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag, 2009. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7. URL: https://www.springer.com/gp/book/9780387848570 (visited on 01/25/2020).

[26] Petter Holme and Beom Jun Kim. "Growing scale-free networks with tunable clustering". In: *Phys. Rev. E* 65 (2 Jan. 2002), p. 026107. DOI: 10.1103/PhysRevE.65.026107. URL: https://link.aps.org/doi/10.1103/PhysRevE.65.026107.

[27] Rebekah L. Horn et al. "Parallel evolution of site specific changes in divergent caribou lineages". In: *Ecology and Evolution* 8.12 (May 2018), pp. 6053–6064. ISSN: 2045-7758. DOI: 10.1002/ece3.4154. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6024114/ (visited on 01/20/2020).

[28] Shaun M. Hug and Brandon S. Gaut. "The phenotypic signature of adaptation to thermal stress in Escherichia coli". In: *BMC Evolutionary Biology* 15.1 (Sept. 2015), p. 177. ISSN: 1471-2148. DOI: 10.1186/s12862-015-0457-3. URL: https://doi.org/10.1186/s12862-015-0457-3 (visited on 01/26/2020).

[29] Iain M. Johnstone. "On the distribution of the largest eigenvalue in principal components analysis". In: *Ann. Statist.* 29.2 (Apr. 2001), pp. 295–327. DOI: 10.1214/aos/1009210544. URL: https://doi.org/10.1214/aos/1009210544.

[30] Pierre Legendre and Loic FJ Legendre. *Numerical ecology*. Vol. 24. Elsevier, 2012.

[31] Richard E. Lenski et al. "Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations". In: *The American Naturalist* 138.6 (1991), pp. 1315–1341. ISSN: 00030147, 15375323. URL: http://www.jstor.org/stable/2462549.

[32] Stephen De Lisle and Daniel I. Bolnick. "A Multivariate View of Parallel Evolution". en. In: *bioRxiv* (Jan. 2020), p. 2020.01.26.920439. DOI: 10.1101/2020.01.26.920439. URL: https://www.biorxiv.org/content/10.1101/2020.01.26.920439v1 (visited on 01/27/2020).

[33] Ulrike von Luxburg. "Clustering Stability: An Overview". English. In: *Foundations and Trends in Machine Learning* 2.3 (Apr. 2010), pp. 235–274. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000008. URL: https://www.nowpublishers.com/article/Details/MAL-008 (visited on 01/25/2020).

[34] Gil McVean. "A Genealogical Interpretation of Principal Components Analysis". en. In: *PLOS Genetics* 5.10 (Oct. 2009), e1000686. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000686. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000686 (visited on 01/14/2020).

[35] Thomas P. Minka. "Automatic Choice of Dimensionality for PCA". In: *Advances in Neural Information Processing Systems 13*. Ed. by T. K. Leen, T. G. Dietterich, and V. Tresp. MIT Press, 2001, pp. 598–604. URL: http://papers.nips.cc/paper/1853-automatic-choice-of-dimensionality-for-pca.pdf (visited on 01/26/2020).

[36] NetworkX developer team. *NetworkX*. 2014. URL: https://networkx.github.io/.

[37] Emmanuel Noutahi. *Fisher's exact test for M x N contingency tables*. 2018. DOI: 10.5281/zenodo.2587757.

[38] W. M. Patefield. "Algorithm AS159. An efficient method of generating r x c tables with given row and column totals". In: *Applied Statistics* (1981), pp. 91–97.

[39] Nick Patterson, Alkes L. Price, and David Reich. "Population Structure and Eigenanalysis". en. In: *PLOS Genetics* 2.12 (Dec. 2006), e190. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.0020190. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020190 (visited on 01/15/2020).

[40] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[41] Barbara Pickersgill. "Parallel vs. Convergent Evolution in Domestication and Diversification of Crops in the Americas". English. In: *Frontiers in Ecology and Evolution* 6 (2018). ISSN: 2296-701X. DOI: 10.3389/fevo.2018.00056. URL: https://www.frontiersin.org/articles/10.3389/fevo.2018.00056/full (visited on 01/20/2020).

[42] S. Joe Qin and Ricardo Dunia. "Determining the number of principal components for best reconstruction". en. In: *Journal of Process Control* 10.2 (Apr. 2000), pp. 245–250. ISSN: 0959-1524. DOI: 10.1016/S0959-1524(99)00043-8. URL: http://www.sciencedirect.com/science/article/pii/S0959152499000438 (visited on 01/26/2020).

[43] Steven Roman. *Advanced Linear Algebra*. Vol. 135. Graduate Texts in Mathematics. New York, NY: Springer New York, 2008. ISBN: 0387728287. DOI: 10.1007/978-0-387-72831-5. URL: http://link.springer.com/10.1007/978-0-387-72831-5 (visited on 01/12/2020).

[44] Jonathon Shlens. "A Tutorial on Principal Component Analysis". In: *arXiv:1404.1100 [cs, stat]* (Apr. 2014). arXiv: 1404.1100. URL: http://arxiv.org/abs/1404.1100 (visited on 01/14/2020).

[45] Nick W. Smith et al. "The Classification and Evolution of Bacterial Cross-Feeding". English. In: *Frontiers in Ecology and Evolution* 7 (2019). ISSN: 2296-701X. DOI: 10.3389/fevo.2019.00153. URL: https://www.frontiersin.org/articles/10.3389/fevo.2019.00153/full (visited on 01/20/2020).

[46] Alexander Soshnikov. "A Note on Universality of the Distribution of the Largest Eigenvalues in Certain Sample Covariance Matrices". In: *Journal of Statistical Physics* 108.5 (2002), pp. 1033–1056. DOI: 10.1023/A:1019739414239. URL: https://doi.org/10.1023/A:1019739414239.

[47] Yoel E. Stuart et al. "Contrasting effects of environment and genetics generate a continuum of parallel evolution". In: *Nature Ecology & Evolution* 1.6 (2017), p. 0158. DOI: 10.1038/s41559-017-0158. URL: https://doi.org/10.1038/s41559-017-0158.

[48] Terence Tao. *Topics in Random Matrix Theory*. English. Providence, R.I: American Mathematical Society, Mar. 2012. ISBN: 978-0-8218-7430-1.

[49] Olivier Tenaillon et al. "Tempo and mode of genome evolution in a 50,000-generation experiment". In: *Nature* 536.7615 (Aug. 2016), pp. 165–170. ISSN: 0028-0836. DOI: 10.1038/nature18959. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4988878/ (visited on 01/21/2020).

[50] Olivier Tenaillon et al. "The Molecular Diversity of Adaptive Convergence". In: *Science* 335.6067 (2012), pp. 457–461. ISSN: 0036-8075. DOI: 10.1126/science.1212986. URL: https://science.sciencemag.org/content/335/6067/457.

[51] Craig A. Tracy and Harold Widom. "Level-spacing distributions and the Airy kernel". In: *Comm. Math. Phys.* 159.1 (1994), pp. 151–174. URL: https://projecteuclid.org:443/euclid.cmp/1104254495.

[52] Caroline B. Turner, Christopher W. Marshall, and Vaughn S. Cooper. "Parallel genetic adaptation across environments differing in mode of growth or resource availability". In: *Evolution Letters* 2.4 (Aug. 2018), pp. 355–367. ISSN: 2056-3744. DOI: 10.1002/evl3.75. (Visited on 01/14/2020).

[53] Timothy M. Wannier et al. "Adaptive evolution of genomically recoded Escherichia coli". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.12 (2018), pp. 3090–3095. ISSN: 1091-6490. DOI: 10.1073/pnas.1715530115.

[54] B. L. Welch. "The Significance of the Difference Between Two Means when the Population Variances are Unequal". In: *Biometrika* 29.3/4 (1938), pp. 350–362. ISSN: 0006-3444. DOI: 10.2307/2332010. URL: https://www.jstor.org/stable/2332010 (visited on 01/28/2020).

[55] Robert Woods et al. "Tests of parallel molecular evolution in a long-term experiment with Escherichia coli". en. In: *Proceedings of the National Academy of Sciences* 103.24 (June 2006), pp. 9107–9112. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0602917103. URL: https://www.pnas.org/content/103/24/9107 (visited on 01/20/2020).

[56] Song Xu, Lucas BÃ¶ttcher, and Tom Chou. "Diversity in Biology: definitions, quantification and models". eng. In: *Physical Biology* (Jan. 2020). ISSN: 1478-3975. DOI: 10.1088/1478-3975/ab6754.

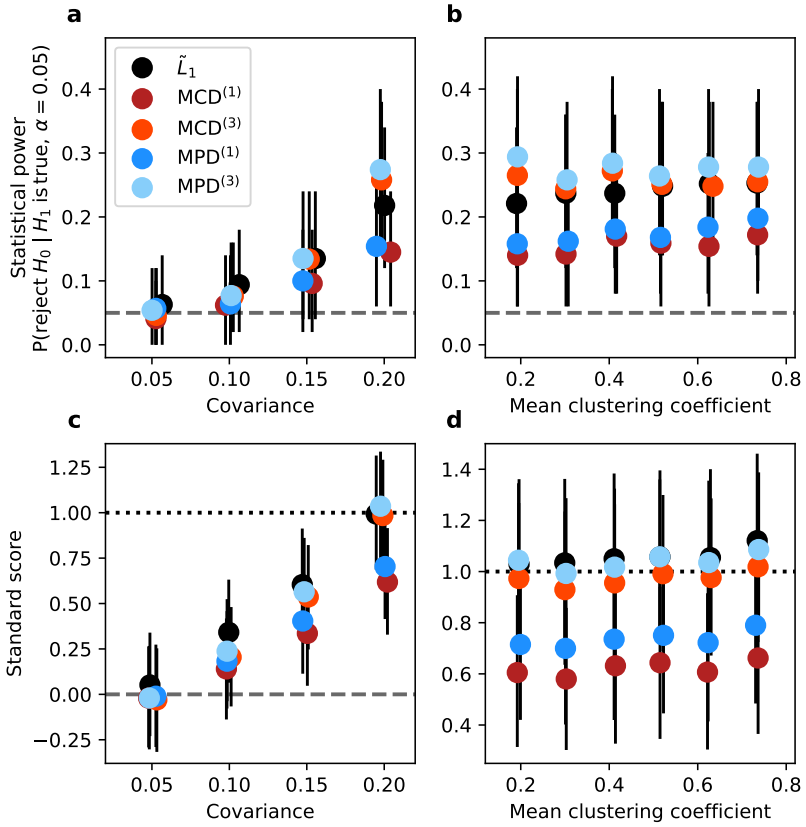

Figure 1: The relationship between properties of $\mathbf{\Sigma}$ and statistical power at a significance level of $\alpha = 0.05$ (dashed horizontal grey line), the probability of rejecting the null hypothesis $\mathbf{\Sigma} = \mathbf{I}$. **a)** Statistical power increases with covariance across all methods, though MCD and MPD only approach the level of $\tilde{L}_1$ when they are estimated over the first three principal components. **b)** There is no clear relationship between statistical power and the degree of clustering in $\mathbf{\Sigma}$. Similar results were found for the standard score of each method in **c)** and **d)**, where the grey and black lines represent values of zero and a single standard deviation, respectively. Power was calculated from 1,000 simulations using 100 replicate populations and 50 genes. Black gray bars represent 95% bootstrapped confidence intervals from 10,000 samples.

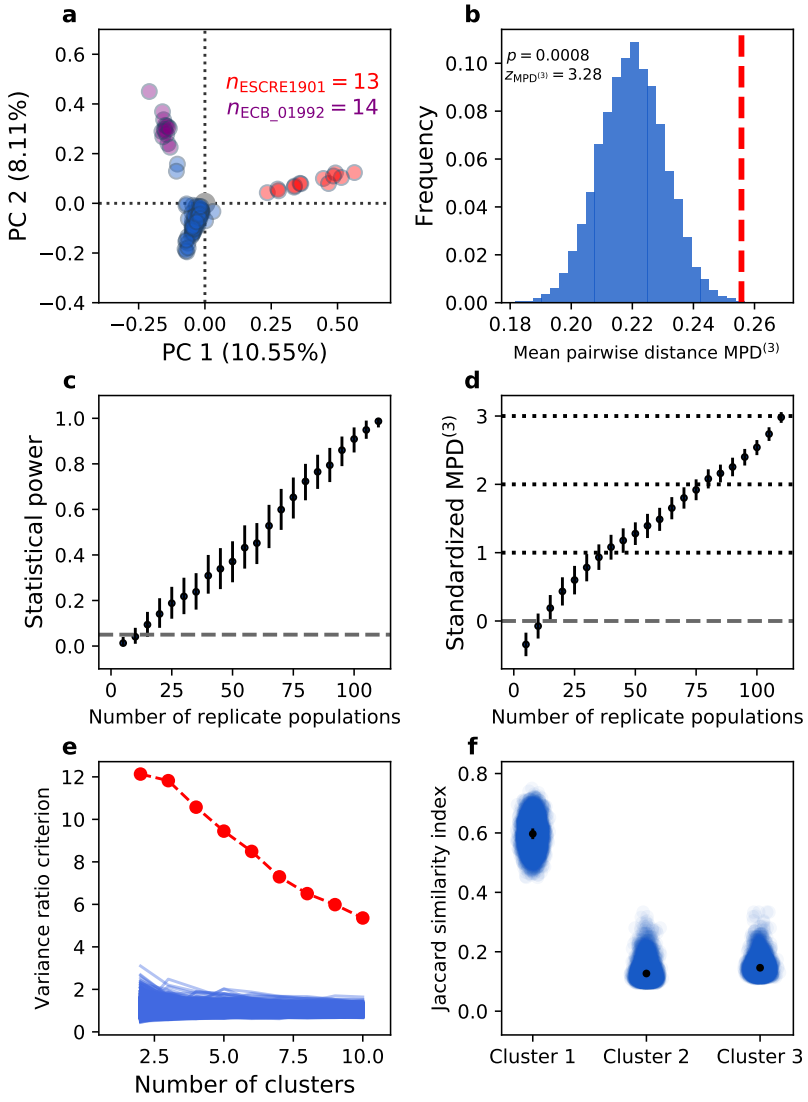

Figure 2: Properties of parallelism in the evolved *E. coli* replicate populations from [50]. **a)** There is clear structure in the data and **b)** MPD$^{(3)}$ (dashed red vertical line) is larger than the null distribution calculated from randomized population-by-gene multiplicity matrices (blue histogram). **c)**, **d)** Covariance is difficult to detect and requires a large number of replicate populations. **e)** While there is clearly greater variance between groups than within, **f)** there is low cluster stability for $k = 3$.

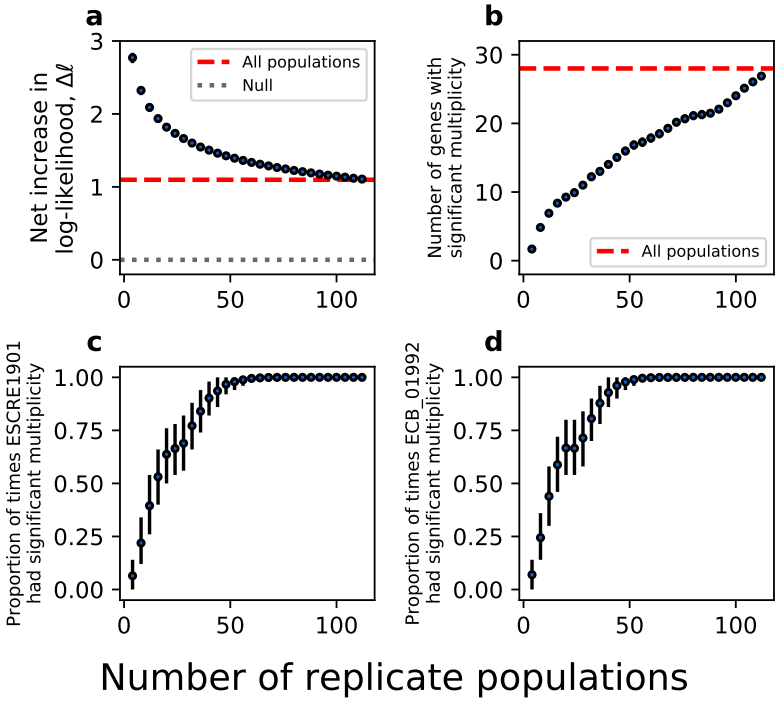

Figure 3: Sampling curve describing how parallelism changes as the number of replicate populations increases using data from [50]. Significant genes in **b)**, **c)**, and **d)** were determined using the multiplicity calculations presented in [23] with a FDR of 0.05. Each dot was calculated from 10,000 sampling events of a given size without replacement from the gene-by-population matrix. Black bars represent 95% bootstrapped confidence intervals calculated from 10,000 samples.

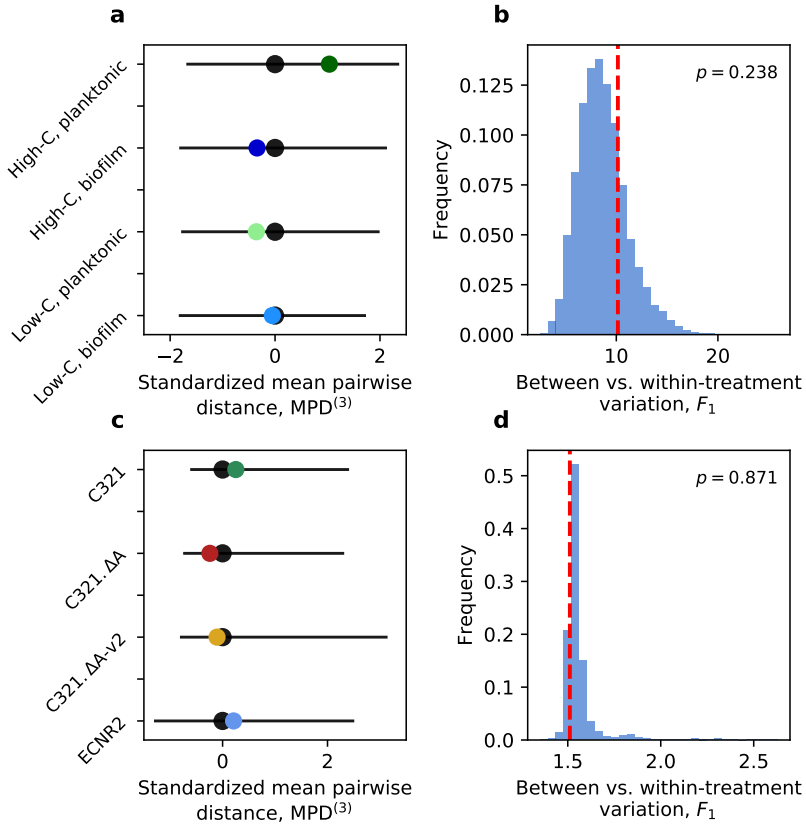

Figure 4: Between covariance is unlikely to affect the degree to detect divergent evolution. $\text{MPD}^{(3)}$ of each treatment and and F_1 statistics across all treatments are not significantly different from the null expectation when covariance between individuals within the same treatment is removed for data from [52] in **a), b)** and data from [53] in **c), d)**. The black dots and lines in **a)** and **c)** represent the mean and 95% standardized CIs from null simulations while the colored dots represent the observed standardized values of $\text{MPD}^{(3)}$. The red dashed vertical lines in **b)** and **d)** represents the observed value of F and the blue histogram represents simulated values of F_1 in the absence of within group covariance.

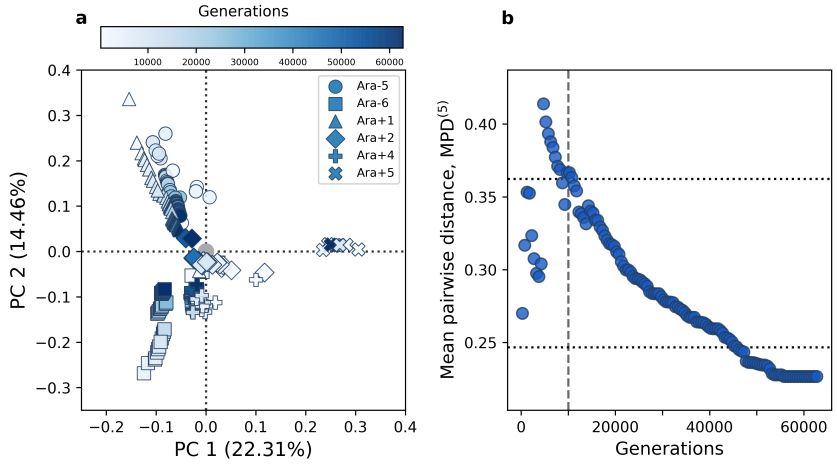

Figure 5: Temporal patterns of parallelism in the LTEE [23]. **a)** The PCA projection of the gene-by-sample multiplicity matrix. **b)** By calculating $\text{MPD}^{(5)}$ at each timepoint we can see temporal patterns in the similarity between populations. The dotted horizontal black lines represent the 95% intervals for MPD in the absence of temporal autocorrelation and the vertical dashed grey line represents the 10,000 generation mark.