Quantifying and Visualizing Parallel Evolution in Replicate Experimental Populations

William R. Shoemaker, Jay T. Lennon

07 April, 2018

1) Background

One of the goals of our research is to determine whether independently evolving populations evolve in similar ways under severe energy-limitation. We have been running our own long-term evolution experiments for several years now and want to be able to quantify the degree of parallel evolution within treatment groups, so we can determine whether different treatment groups are evolutionarily diverged. However, the degree of parallelism is often not quantified in experimental evolution, the degree of expected parallelism depends on the unit of interest (i.e., gene, trait, fitness), and the set of available statistical and visual tools is fairly limited. Recent attempts to quantify parallel evolution focused on the excess number of substitutions in a gene relative to a null model (Tenaillon et al., 2016; Good et al., 2017). This approach is innovative and can likely be applied to pooled sequencing data from any evolution experiment that has been run on a long enough time-scale that a large number of mutation origin-fixation events are likely to have occurred. It is worth noting the theoretical implications of this approach. That is, focusing on parallelism at the gene level implies that genotypic space (i.e., the set of all possible genotypic combinations) has been coarse-grained from the nucleotide to the gene level, where the value at a node is now the number of mutations observed in the set of sites within the gene rather than the presence or absence of a mutation at a single site. If the set of evolutionary paths to a fitness optimum (assuming selection is sufficiently strong, and mutation is sufficiently weak) include sites in different genes that have a negative effect on fitness in combination (site epistasis), replicate populations may be enriched for mutations in different sets of genes.

To accomplish our goal of quantifying the degree of parallel evolution and visualizing evolutionary trajectories in experimental populations, we are taking a three-step approach:

1) Apply and developing a multivariate statistical approach to quantify and visualize evolutionary trajectories from pooled population sequence data.

2) Confirm the validity of our approach through simulated data of evolutionary dynamics.

3) Test our approach on an existing high-quality dataset.

In order to determine the degree that independently evolving replicate populations arrive at the same evolutionary outcome, we initially extended the G-score measure presented in Tenaillon et al. (2016) to the level of individual populations (rather than being a measure that aggregates across replicate populations) and applied it to our long-term evolution experiment using multivariate ordination techniques (Shoemaker and Lennon, 2017).

Since then, a more thorough attempt at quantifying parallel evolution was done in a recent study that examined fine-scale temporal pooled sequencing data from Richard Lenski's Long-term Evolution Experiment (Good et al., 2017). We have since worked to apply the measures and statistics presented in Good et al. to our multivariate approach (2017).

As a brief review, in Good et al. (2017) the authors propose a measure of gene multiplicity to detect evolutionary parallelism at the gene level among replicate populations. The multiplicity for each gene is

$$m_i = n_i \cdot \frac{\overline{L}}{L_i}$$

where n_i is the number of mutations in gene i across all replicate populations, L_i is the number of non-synonymous sites in gene i, and $\overline{L}$ is the average value of L_i across all genes in the genome. Under the null model that the probability that a gene contains a mutation is simply proportional to the length of the gene ($p_i \propto L_i$), all genes have the same expected multiplicity $\overline{m} = n_{tot}/n_{genes}$, where n_{tot} is the number of mutations among all replicate populations and n_{genes} is the number of genes in the genome.

In Good et al. (2017) the authors determine that in nonmutator LTEE populations approximately half of all mutations occurred in genes with $m_i \geq 2$, twice as many as expected under the null model. The authors concluded that the null model should be replaced with an alternative where mutations are assigned to each gene with probability

$$p_i \propto L_i r_i$$

where r_i is an enrichment factor that is not equal to 1. Under the alternative model the maximum likelihood estimator for the enrichment factor is the ratio of observed and expected multiplicities, $r_i = m_i/\overline{m}$ and the net increase relative to the null model across all genes is

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

The authors note that the maximum likelihood estimate r_i may overfit the data and propose that a more appropriate alternative model would be one that focuses on a subset I of the genes where $r_i \neq 1$, while the remaining genes have $r_i = 1$. The authors identify this set of genes using a cricial P-value, P^*, for a the False Discovery Rate $\alpha = 0.05$ and modify the enrichment factors as follows

$$r_i = \left\{ \begin{array}{cc} \frac{m_i}{\overline{m}} \left(\frac{1 - \frac{\sum_{i \epsilon I} L_i}{\overline{L} n_{genes}}}{1 - \frac{\sum_{i \epsilon I} n_i}{n_{tot}}} \right) & \text{if } i \epsilon I \\ 1 & \text{else.} \end{array} \right\}$$

This is an innovative approach that builds off of statistical distributions used to describe parallel evolutionary outcomes. However, this measure pools the mutation data for all replicate populations for each gene. To allow for the comparison between replicate populations so that we can begin to develop statistics to determine whether replicate populations have similar evolutionary trajectories from pooled sequencing, the multiplicity statistics presented in Good et al. (2017) need to be deconstructed to the level of individual populations. To accomplish this goal, we propose a multiplicity measure for the ith gene in the jth population

$$m_{i,j} = n_{i,j} \cdot \frac{\overline{L}}{L_i}$$

with the expected multiplicity in population j of $\overline{m_j} = n_{tot,j}/n_{genes}$, giving a log-likelihood compared to the null model (which is now $r_{i,j} = m_{i,j}/\overline{m_j}$)

$$\Delta\ell_j = \sum_i n_{i,j} \log\left(\frac{m_{i,j}}{\overline{m_j}}\right)$$

and the modified enrichment factor

$$r_{i,j} = \left\{ \begin{array}{cc} \frac{m_{i,j}}{\overline{m_j}} \left(\frac{1 - \frac{\sum_{i \epsilon I} L_i}{\overline{L} n_{genes}}}{1 - \frac{\sum_{i \epsilon I} n_{i,j}}{n_{tot,j}}} \right) & \text{if } i \epsilon I \\ 1 & \text{else.} \end{array} \right\}$$

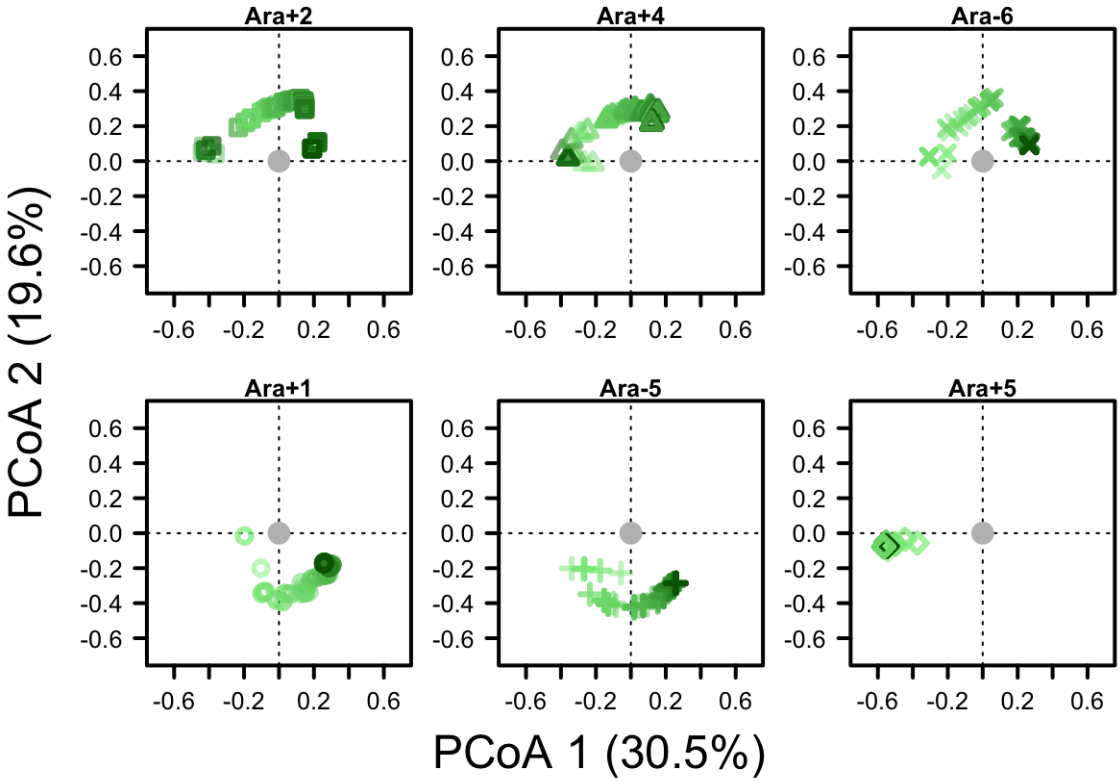

Figure 1: The first two axes of a PCoA on the LTEE gene-by-population multiplicity matrix. Each population is plotted as a separate figure. Lighter colored points indicate earlier timepoints while darker timepoints indicate later timepoints. The grey dot indicates the ancestor (i.e., no mutations in any gene)

Using these modified population level gene enrichment scores and the publically available data presented in Good et al. (2017), we calculate the multiplicity score for each gene within each population at each time point for all genes within set I, generating a gene-by-population multiplicity matrix. We then built a Bray-Curtis dissimilarity matrix and used Principal Coordinates Analysis (PCoA) to reduce the dimensionality of the dataset and visualize the evolutionary trajectories of the six nonmutator populations. The code and additional analyses can be found in the GitHub repository.

2) Visualizing evolutionary trajectories through coarse-grained genotypic space

```
pdf
  2
```

We see from the ordination (Fig. 1) that Ara+2, Ara+4, and Ara-6 have similar non-linear trends in ordination space. Likewise, Ara+1 and Ara-5 have similar trends. However, Ara+5 shows a divergent trend and we do not have an immediate explanation. To determine the sets of genes that contribute to each of these trajectories as well as how the variation in signatures of parallelism changes over time, we will be adapting appropriate measures presented in Good et al. (eqs. 80 - 87 in the supplement) to account for variation between populations.

We can make this trend clearer by plotting the first two axes as a function of time.

```
pdf
```

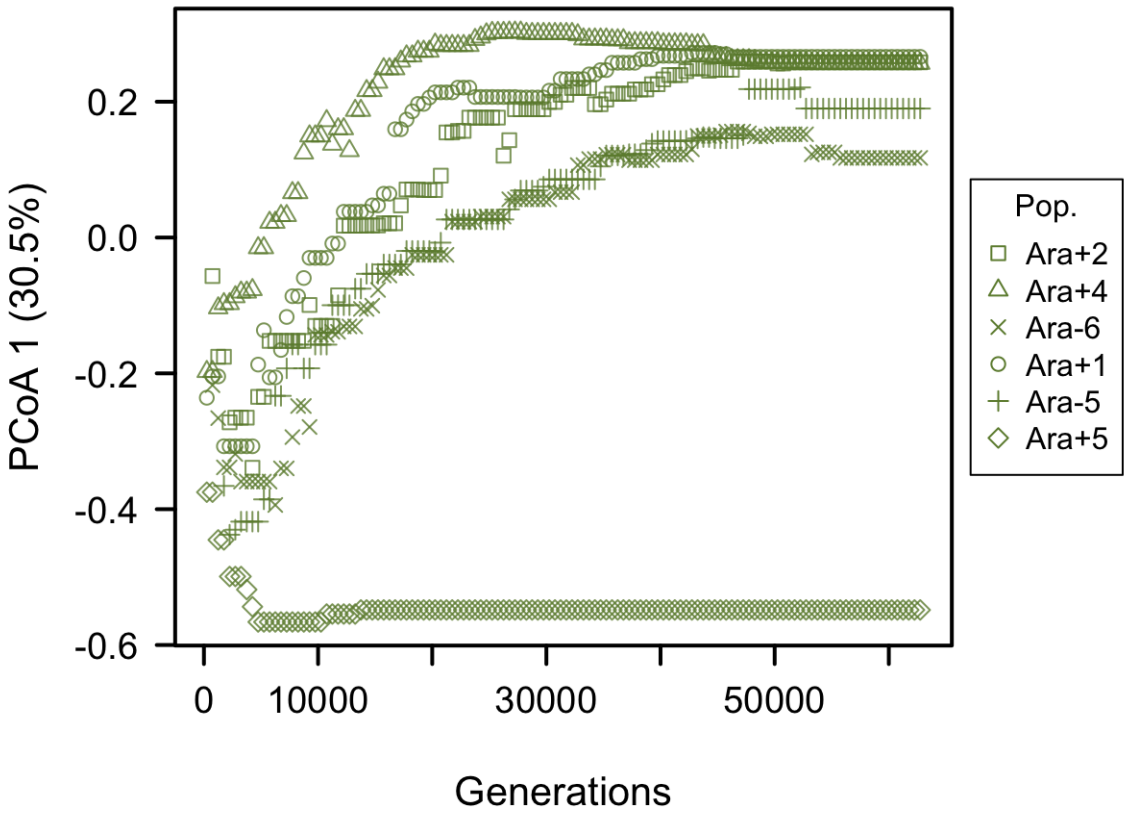

Figure 2: The first axis of the PCoA axis plotted against time (generations).

```
2
```

The trajectory of Ara+5 remains fairly constant over time. This may be because Ara+5 is not accumulating mutations in the set of genes that passed false-discovery rate correction (I). We will be examining this in subsequent analyses and will conduct the same PCoA analyses on the gene-by-population multiplicity matrix of all protein encoding genes, rather than just the set of genes within set I.

For the remaining five nonmutator populations, the trajectory of the major axis of the ordination over time bears a strikng resemblance to the trajectory of fitness gains. Good and Desai (2015) previously determined that a decelerating fitness trajectory by itself provides little power to distinguish between evolutionary models models. Including the mutation trajectory (i.e., the number of mutations per-clone, $(\partial_t \overline{M}(t))$) provided strong constraints on the set of models that explain the fitness trajectory. Perhaps examining the mutation trajectories of individual genes (i.e., a vector of $\partial_t \overline{M_i}(t)$ for all i rather than $\partial_t \overline{M}(t)$) would further constrain the set of models that explain the fitness trajectory.

We can examine the degree divergence in the five nonmutator populations (excluding Ara+5) by plotting the second PCoA axis against the number of generations.

```
## pdf
##   2
```

We can see that on the second PCoA axis the trajectories of Ara+1, Ara-5, and Ara-6 overlap while Ara+2 and Ara+4 are fairly close. Again, the trajectory of Ara+5 remains fairly constant over time.

We are working to apply time-series clustering techniques on the ordination results to explore how multiple populations can be grouped as a single trajectory.

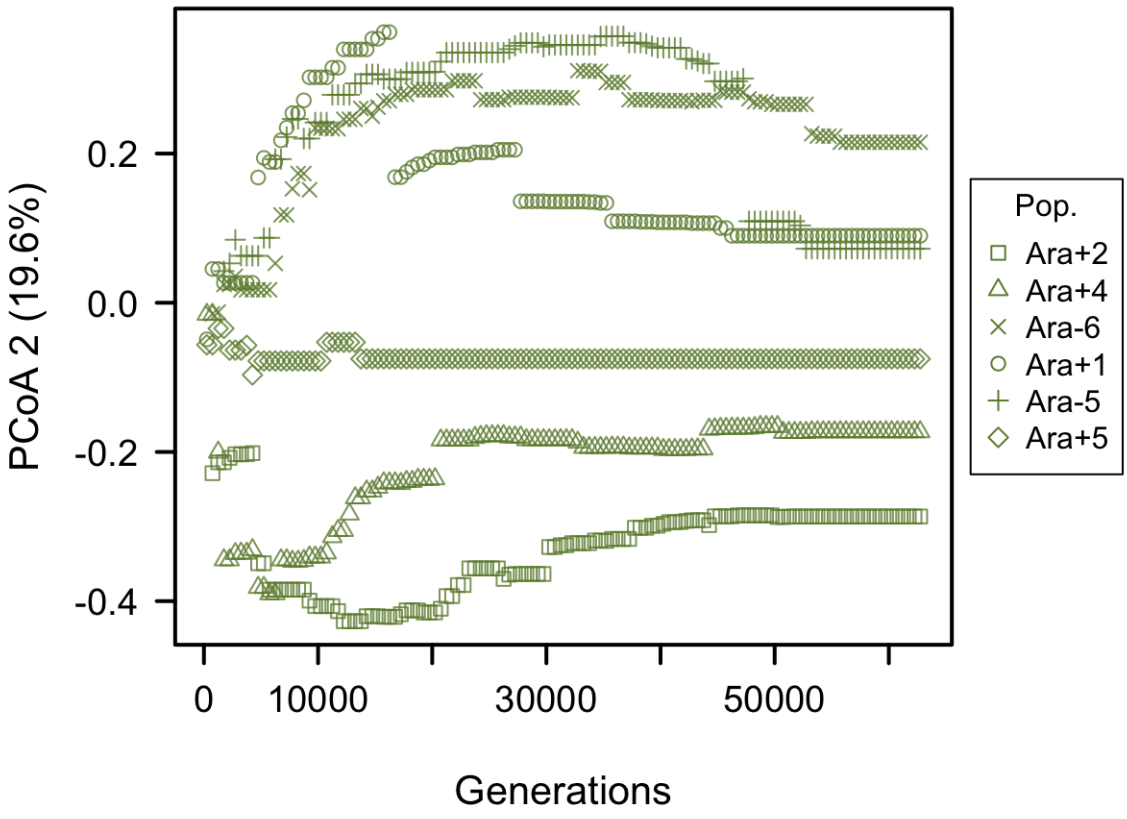

Figure 3: The second axis of the PCoA axis plotted against time (generations).

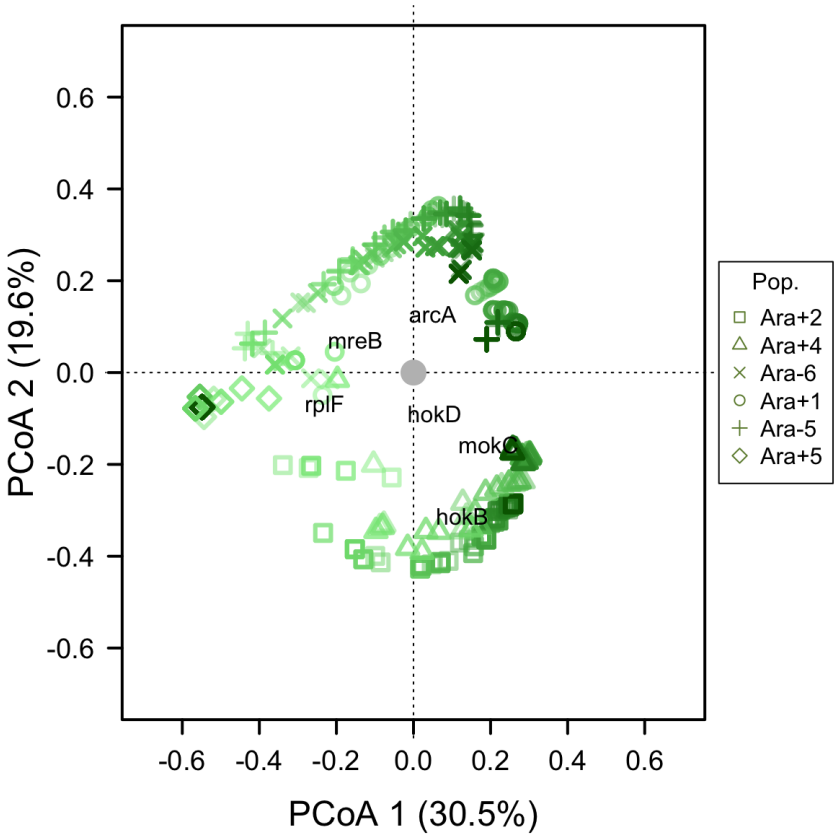

Figure 4: The same PCoA as shown in Fig. 1, but with gene coordinates plotted in ordination space.

We can identify the set of genes contributing to our observed patterns by plotting explanatory vectors (i.e., gene coordinates) in ordination space. These coordinates reflect the strength and direction that each gene has on the ordination of the different time points and allow us to identify what genes contribute o divergent trajectories in ordination space.

```
## pdf
##   2
```

From this visualization, we can see that the *mokC*, *hokB*, and *hokD* genes are close to the later timepoints for the Ara+2 and Ara+4. Whereas *rplF*, *mreB*, and *arcA* tend to be associated with populations Ara-6, Ara+5, and Ara+6. The genes *mokC*, *hokB*, and *hokD* all seem to be part of the type I toxin-antitoxin system, while *rplF*, *mreB*, and *arcA* appear to be part of different molecular pathways.

We can confirm these visual patterns in a more quantitative way by determining the correlation coefficient of each gene along one or PCoA axes, identifying a correlation-coefficient cutoff, and conducting a permutation test.

3) Quantifying variation in evolutionary outcomes

A central question we want to answer from our own evolution experiments is how the degree of similarity between evolving populations changes with time. We can roughly see that the populations reach similar points in Euclidean space over time, but we're only looking at a single ordination axis at a time.

To quantify the degree of similarity between populations over time across multiple ordination axes, we

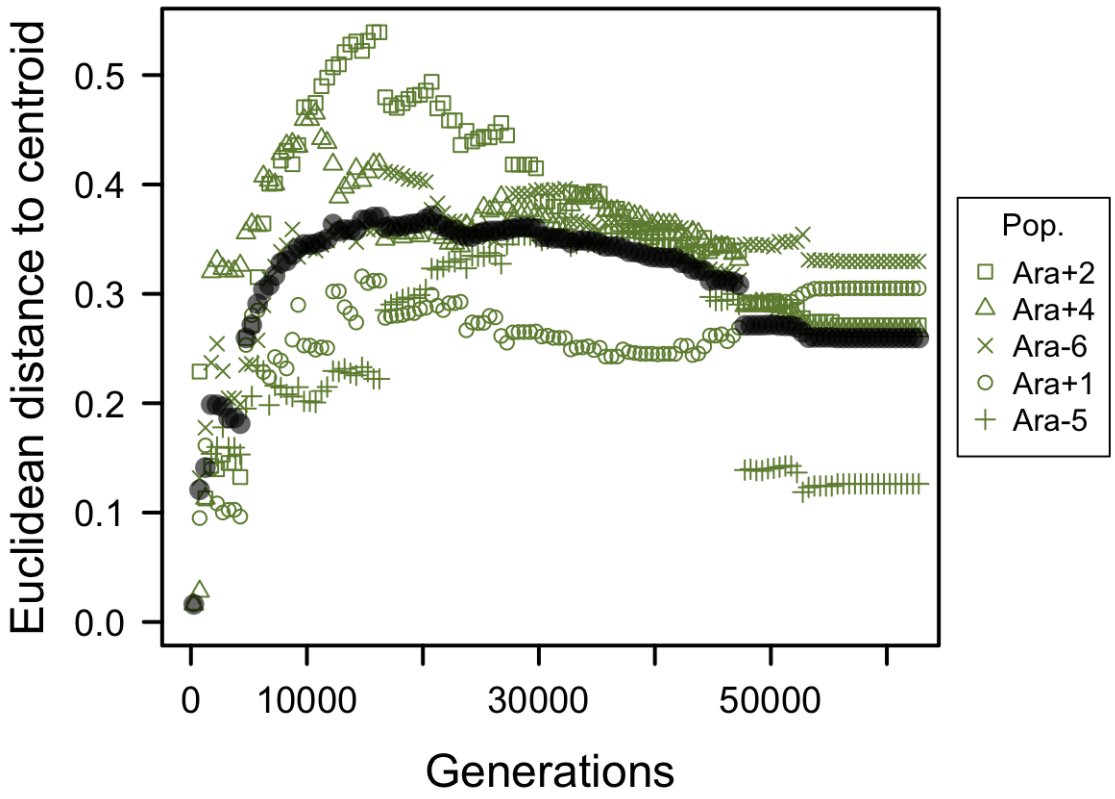

Figure 5: The Euclidean distance to the centroid of all six nonmutator populations plotted against time. The black circles represent mean centroid distance

quantified the Euclidean distance between each population and the median for the first three axes of ordination space for all populations sampled at a given time-point (i.e., the centroid). We then plotted the centroid distance for all populations (green points) and the mean centroid distance (black closed circles) against the number of generations. We chose to exclude the population Ara+5 from this analysis since its trajectory was highly divergent from the other nonmutator populations in the first two axes.

```
## pdf
##   2
```

The scatter of the points around the mean clearly increases for the first 20,000 generation and then continues to decreases, suggesting that the degree of variance in evolutionary outcomes may be a function of time.

We have written code to determine whether or not this result is significant by generating random gene-by-population multiplicity matrices where the probability that a mutation lands on the gene is simply the gene length ($p_i \propto L_i$) and are planning on running a permutation test on the gene-by-population multiplicity matrix.

4) Planned analyses

1) Extending the historical contingency analysis in Good et al. (2017) to the multivariate case.

2) Applying temporal clustering to the PCoA ordination as a means of providing an exploratory analysis of the number of evolutionary trajectories.

3) Determine whether our approach amends itself to time-series analysis by applying certain principles (i.e., temporal autocorrelation) to our analysis.

4) Determine if we get comparable results using alternative transformations and ordination techniques (i.e., Hellinger transformed distance matrix and Principal Component Analysis (PCA)).

5) Extend our analysis to the polymorphism data, which we can treat as the temporal turnover of genetic diversity (analogous to Beta-diversity).

5) Replicate observed patterns via simulation

We are currently working to build off of and simulate existing population genetic models to confirm our observed patterns. We are aiming to take a two-prong approach towards deriving and simulating evolutionary models of parallel evolution:

1) "Bottom-up": For this approach we will simulate genotypic spaces under different evolutionary scenarios (i.e., sign epistasis, historical contingency, etc.), which will then be coarse-grained into individual genes. We've been paying particular attention to the block model (Perelson and Macken, 1995; Schmiegelt and Krug, 2014).

2) "Top-down": Here the degree of parallel evolution is constrained by the genetic background dependence of the distribution of fitness effects. This form of epistasis has been referred to as "macroscopic epistasis" and has been used in a previous approach towards inferring the evolutionary dynamics of the LTEE (Good and Desai, 2015). This approach has the potential advantage of providing a general constraint on the set of evolutionary outcomes based on few pieces of information, since it does not include information about the background dependence of individual mutations.

Our primary aim is to simulate data from different theoretical approaches under various evolutionary dynamics to test the limits of an ordination approach. We are also interested in analytic derivations to describe the degree that parallel outcomes can be observed by methods that coarse-grain genotypic space to the gene level (i.e., the multiplicity score) and are open to exploring the degree of parallelism under alternative regimes (i.e., clonal interference).

6) References

Good, B. H., and M. M. Desai. 2015. The impact of macroscopic epistasis on long-term evolutionary dynamics. Genetics 114:172460

Good, B. H., M. J. McDonald, J. E. Barrick, R. E. Lenski, and M. M. Desai. 2017. The dynamics of molecular evolution over 60,000 generations. Nature 551:45–50

Perelson, A.S., C. A. Macken. 1995. Protein evolution on partially correlated landscapes. Proc. Natl. Acad. Sci. USA 92:8657–9661

Schmiegelt. B., and J. Krug. 2014. Evolutionary accessibility of modular fitness landscapes. Journal of Statistical Physics 154:334–355

Shoemaker, W. R., and J. T. Lennon. 2017. The contribution of dormancy to microbial evolution. Society for Molecular Biology and Evolution, Austin, Texas, USA

Tenaillon, O., et al. 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature 536:165–170