# Evolutionary trajectories in the Long-term Evolution Experiment

*William R. Shoemaker, Jay T. Lennon*

*30 March, 2018*

## 1) Background

One of the goals of our research is to determine whether independently evolving populations are evolving under similar trajectories. One of the more recent and thorough attempts at addressing this goal is a recent study that examined fine-scale temporal sampling of Richard Lenski's Long-term Evolution Experiment (Good et al., 2017). 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

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

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 $i$th gene in the $j$th 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

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

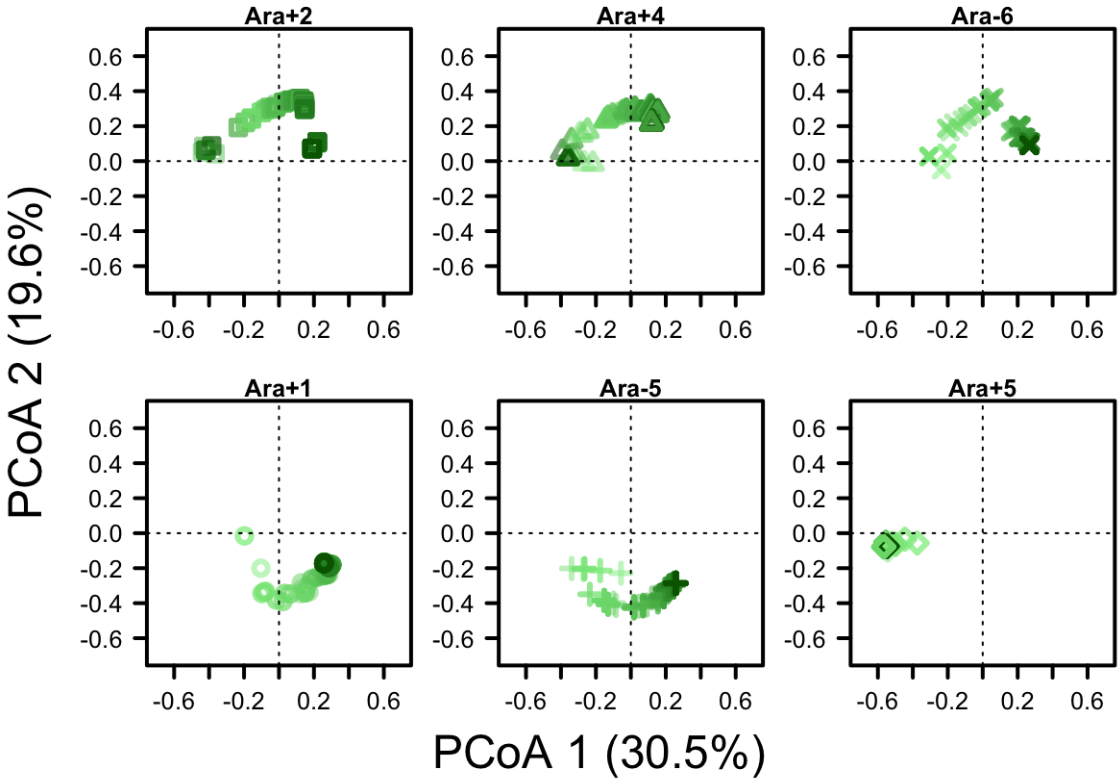
Using these modified population level gene enrichment scores and the publically available data presented in Good et al. (2017), we calculate the multiplicity score for each gene within each population at each time point for all genes within set $I$, generating a gene-by-population multiplicity matrix. 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. To determine whether or not the rate that mutations are acquired in each gene decreases, we calculated the Euclidean distance of the first three PCoA axes between timepoints for each population.

## 2) Setup work environment

## 3) Load and clean data

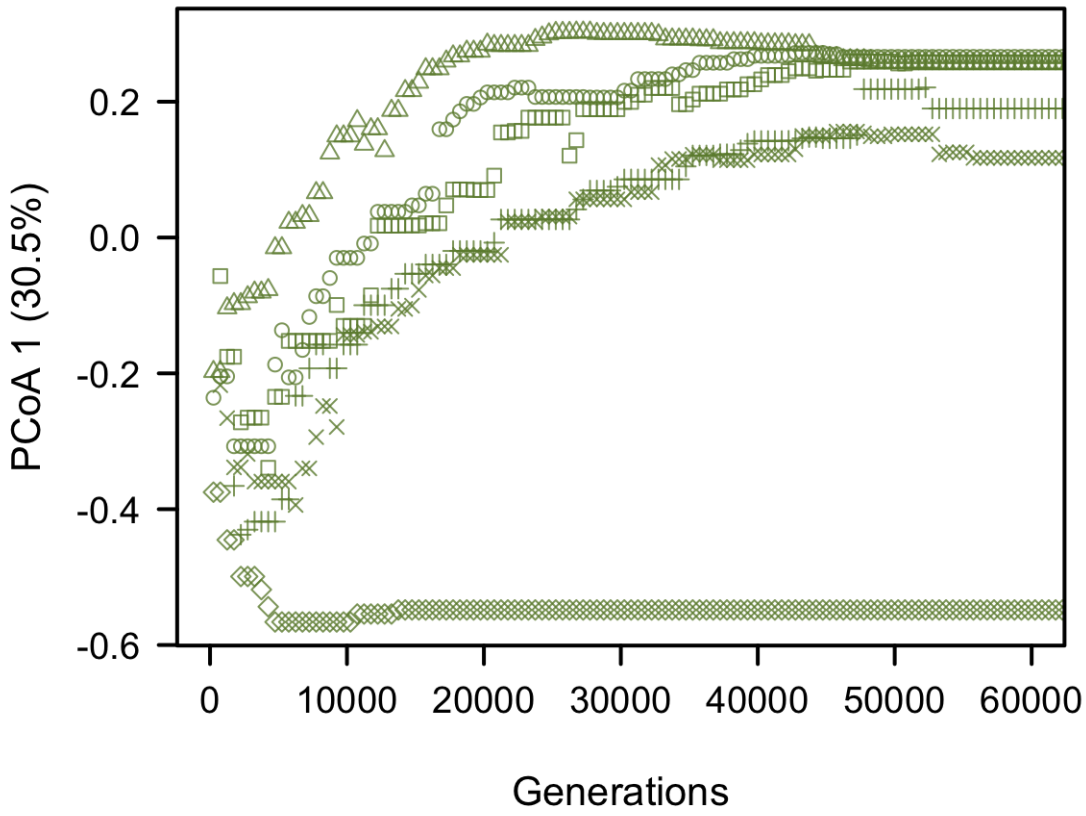
## 4) Make and visualize PCoA ordination

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

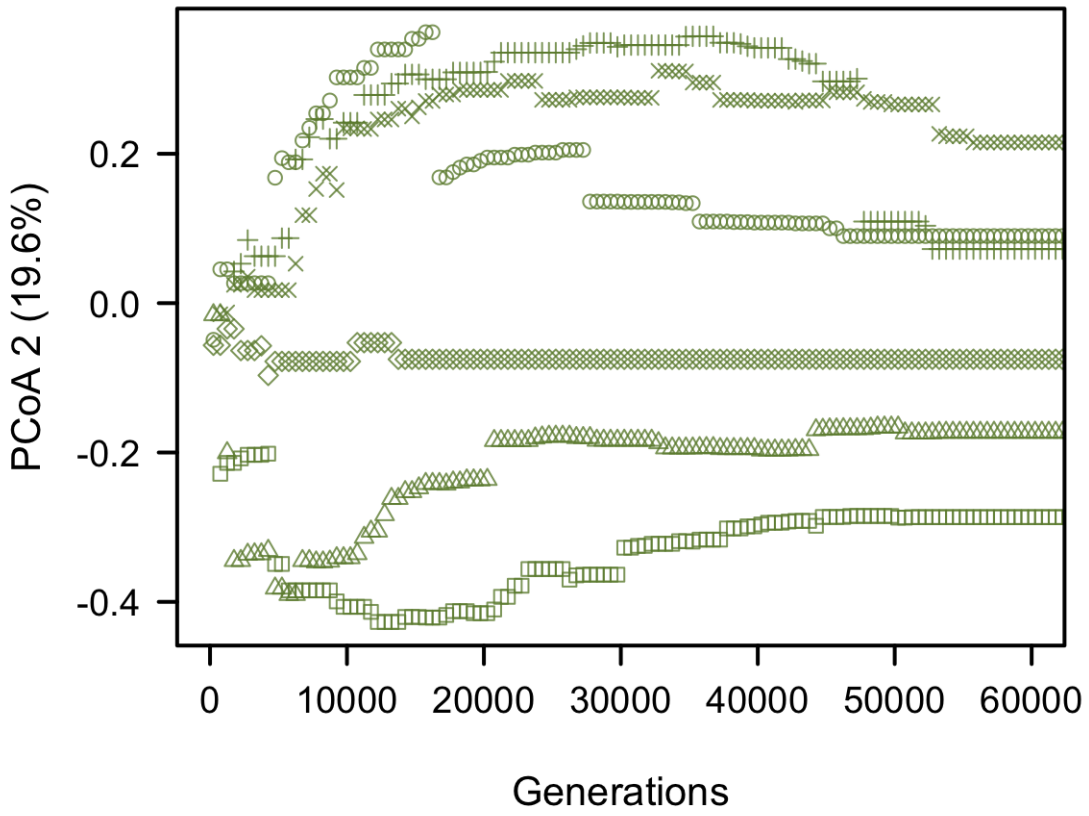



We can make this trend clearer by plotting the first two axes as a function of time.

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


PCoA 1 (30.5%)
0.2
0.0
-0.2
-0.4
-0.6
0
10000
20000
30000
40000
50000
60000
Generations

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




So we see from the ordination 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 does not show much of a trend and there is no 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 are working to apply time-series clustering techniques on the ordination results to explore how multiple populations can be grouped as a single trajectory.

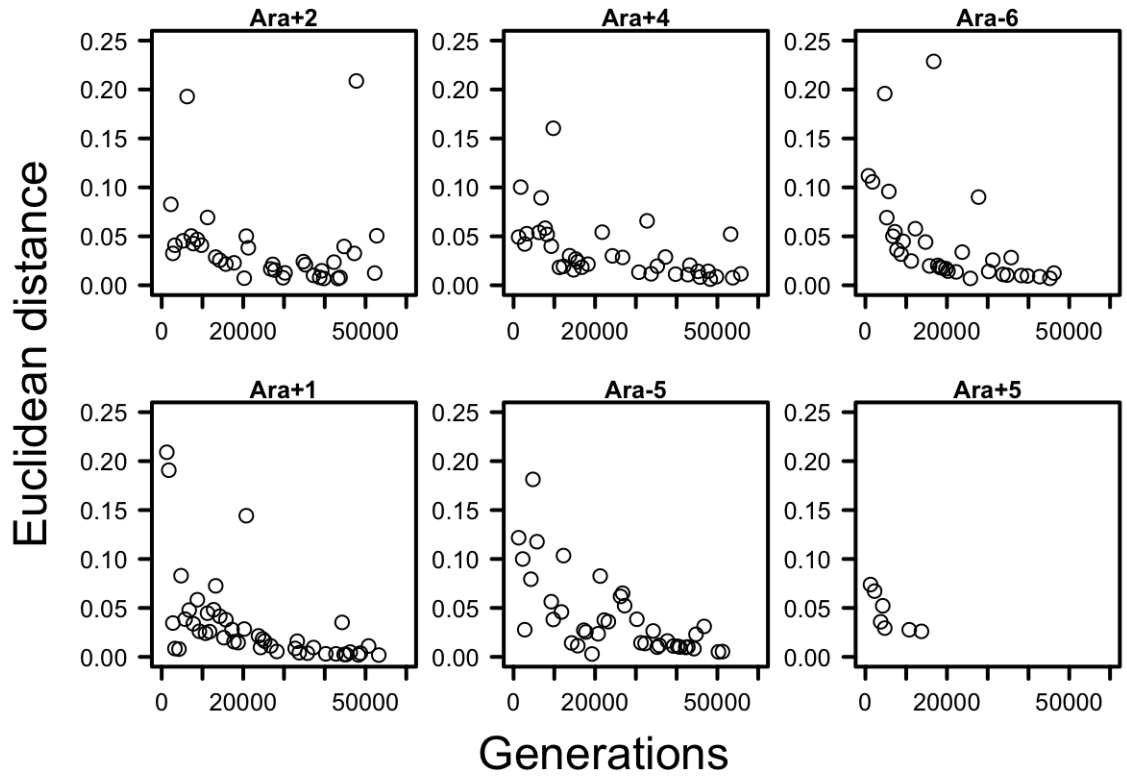
## 5) Quantifing Euclidean distance between timepoints

A basic prediction from Fisher's geometric model (FGM) of adaptive phenotypic evolution is that the magnitude of change in phenotypic space declines over the course of the adaptive walk towards an optimum. While we are using genetic data instead of phenotypic data, it is possible that the rate of change in ordination space could exhibit a qualitatively similar trend. This prediction assumes that some form of FGM describes phenotypic evolution in the LTEE and that the genotype-to-phenotype map is linear to the extent that qualitative similarities between the two distributions are not erased. There are other explanations for this pattern, one of which may be a variation of the global fitness-mediated epistasis model described in Good and Desai (2014).

An example of an adaptive walk towards an optimum on a 2 dimensional form of Fisher's geometric model.

To examine the decay in the rate of change with time, we quantified the pair-wise Euclidean distance ($d()$) between sequential timepoints for the first three PCoA axes ($d(\mathbf{t_i}, \mathbf{t_{i+1}})$) for each population. We then plotted the second time point of the sequential pair-wise distance ($t_{i+1}$) against the Euclidean distance.

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

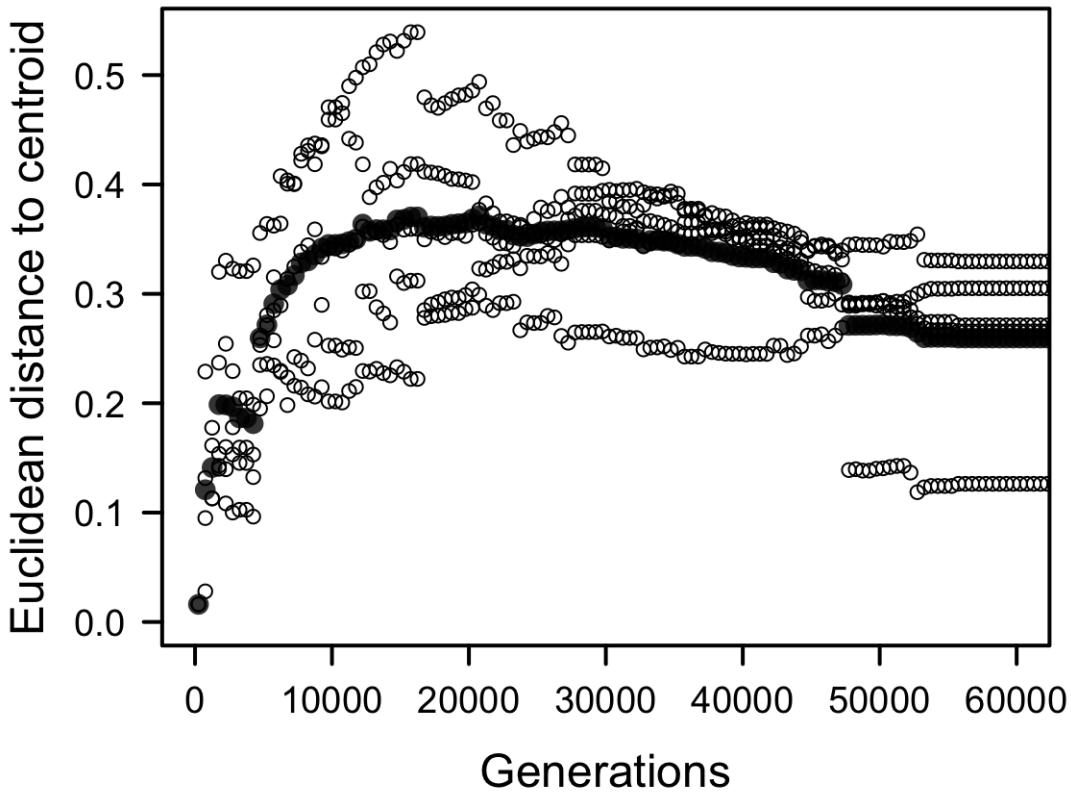



While we see considerable variation, a number of populations show a qualitative trend where Euclidean distance between time points decreases with time. We are working on identifying and applying appropriate statistical models to describe this trend and determine the extent that it's consistent across populations. We are also working to identify the set of potential explanations for this trend.

## 6) Mean centroid distance

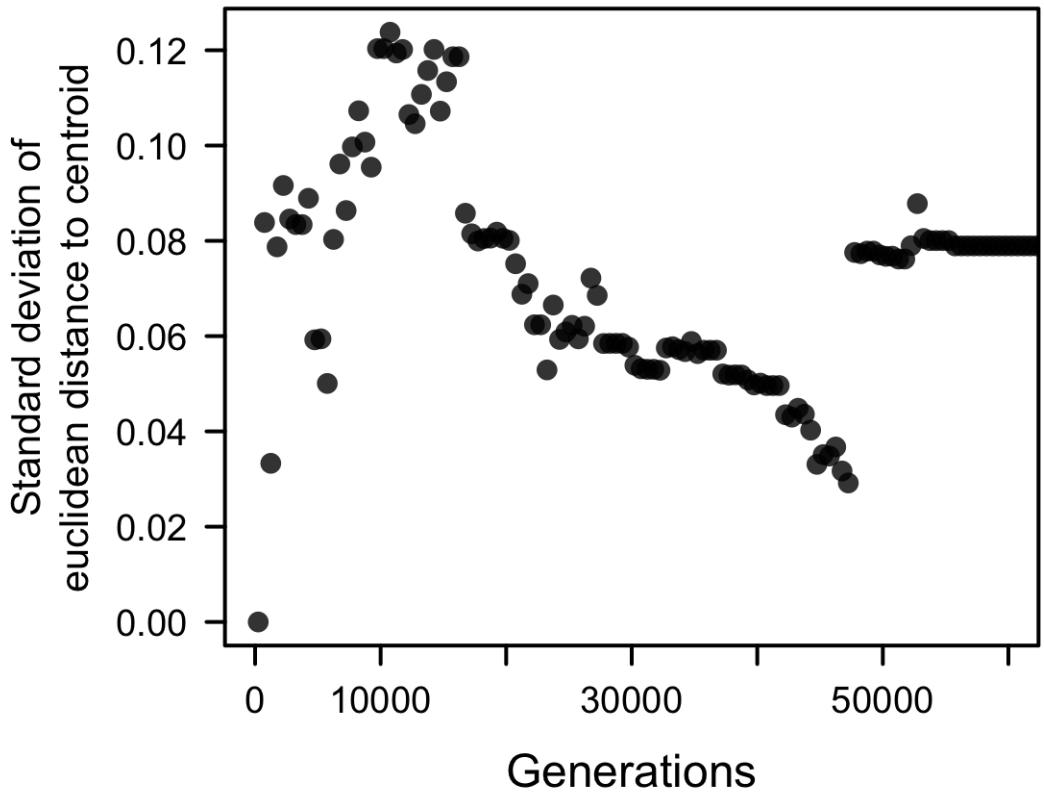
A central question we want to answer from our own evolution experiments is whether independently evolving populations reach similar evolutionary endpoints. To examine this, we quantified the Euclidean distance between each population and the median in ordination space of all populations sampled at a given time-point (i.e., the centroid). We then plotted the centroid distance for all populations (open circles) and the mean centroid distance (closed circles) against generations.

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



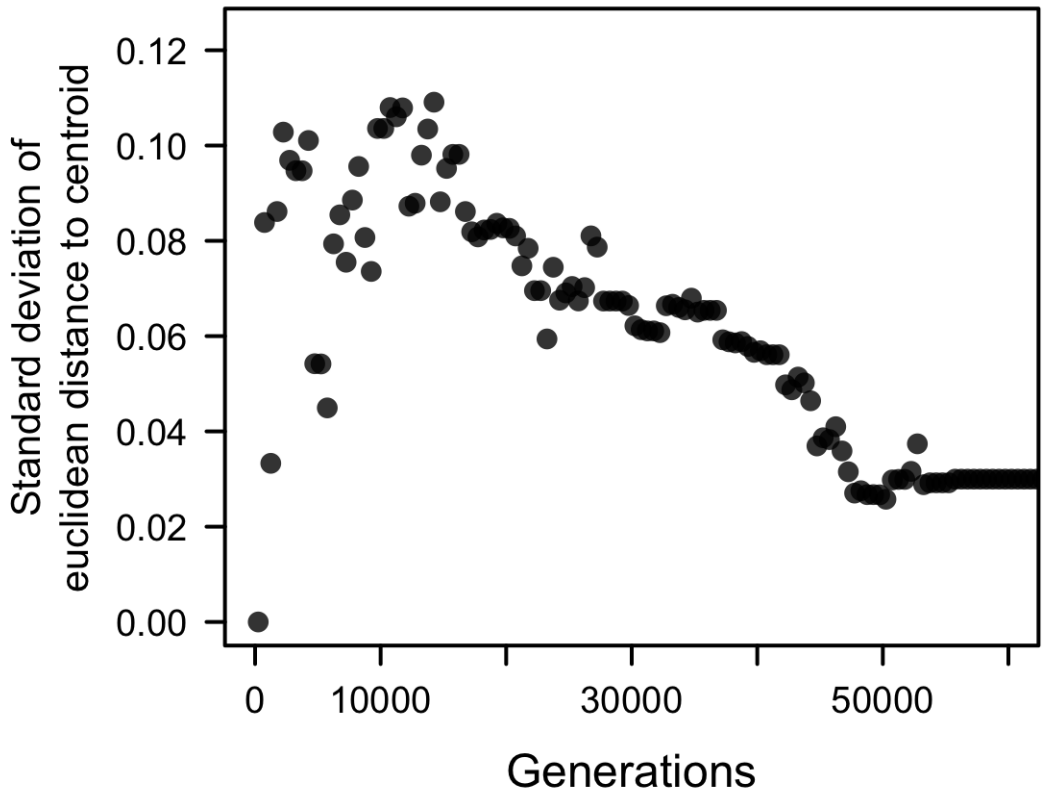

The scatter of the points around the mean clearly decreases with time, suggesting that there may be a relationship between the variance in the degree of parallel evolution with time. To examine this, we plotted the standard deviation of the mean centroid distance against the number of generations.

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




The standard deviation for the last approximately 10,000 generations is clearly disproportionately skewed by some outlier timepoints from population Ara+2, the reason for these unusually high Euclidean distance values in Ara+2 is unclear. If we remove these points and make the same plot, we see that the relationship between the standard deviation of Euclidean distance to the centroid and time is clearly unimodal and right-skewed.

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




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.

## 7) Replicate observed patterns via simulation

We are currently working to build off of and simulate existing population genetic models to confirm our observed patterns. Right now, we're working to simulate a gene-by-population matrix for the running out of mutations model and a global epistasis model (based off of the one proposed in Good and Desai (2014)). Because we observe two different evolutionary trajectories that ultimately reach the same region of PCoA space, sign epistasis could serve as a potential explanation and we are working to include it in our ongoing simulations.

## 8) References

Good, B. H., and M. M. Desai. 2014. 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