

August 19, 2019

1 Introduction

2 Results

We measure two types of covariances between allele frequency caused by selection: temporal autocovariances, and across-replicate covariances. First, positive temporal autocovariance in a neutral allele frequency’s trajectory occurs when the allele becomes associated with a high or low fitness chromosomal background, and this association persists through the generations due to linkage disequilibrium. As long as the direction of selection remains constant, the fitness background is predictive of the direction in the neutral allele’s frequency changes, creating positive covariance. Even though these magnitude of frequency changes at each site may be subtle, as they would be under polygenic selection, cumulatively these perturb neighboring sites in a predictable manner and build up temporal autocovariance which acts as a genome-wide signal of linked selection. Second, if evolution occurs in replicate populations undergoing convergent selection pressure, neutral sites linked to fitness backgrounds shared across replicates are expected to change in the same direction. This creates across-replicate covariance, which is a measure of the extent to which convergent selection pressures across replicate populations cause similar allele frequency changes. Finally, it is important to note that under the null model where the fitness differences between individuals are entirely random and non-heritable (e.g. when selection is not acting), both forms of covariances are expected to be zero.

Explain G.

We first analyzed Barghi et al. (2019), an evolve-and-resequence study with ten replicate populations exposed to a high temperature environment and evolved for 60 generations.

3 Appendix

4 Sampling Bias Corrections

Following Waples (1989), we have that the variance in the initial generation, which is entirely due to the binomial sampling process, is $\text{Var}(p_0) = p_0(1-p_0)/d_0$ where d_0 is the number of binomial draws (e.g. read depth). At a later timepoint, the variance in allele frequency is a result of both the binomial sampling process at time t and the evolutionary process.

Using the law of total variation,

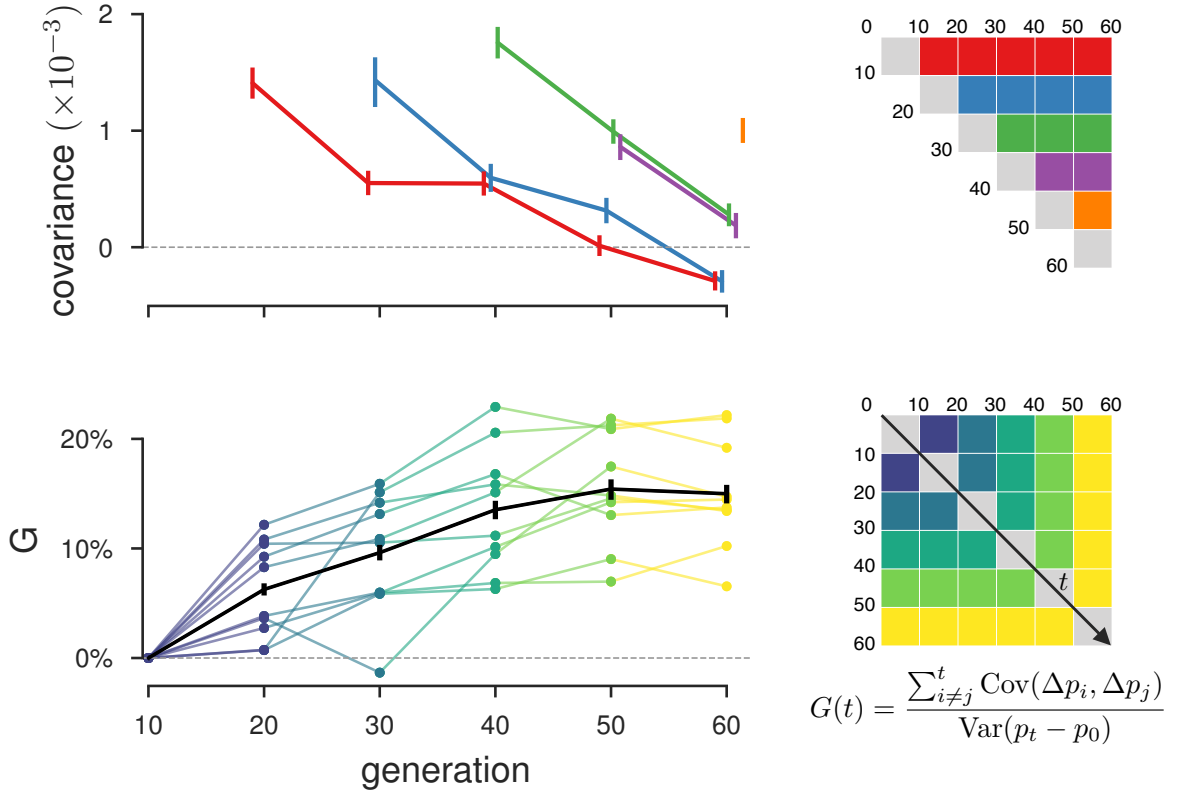


Figure 1: $B = 5000$ bootstraps, with 100e

Figure 2

Figure 3

Figure 4

$$\text{Var}(\tilde{p}_t) = \mathbb{E}(\text{Var}(\tilde{p}_t|p_t)) + \text{Var}(\mathbb{E}(\tilde{p}_t|p_t)) \quad (1)$$

$$= \underbrace{\frac{p_t(1-p_t)}{d_t}}_{\text{generation } t \text{ sampling noise}} + \underbrace{\text{Var}(p_t)}_{\text{variance due to evolutionary process}} \quad (2)$$

Under a drift-only process, $\text{Var}(p_t) = p_0(1-p_0) \left[1 - \left(1 - \frac{1}{2N}\right)^t\right]$. However, with heritable variation in fitness, we need to consider the covariance in allele frequency changes across generations (Buffalo and Coop 2019). We can write

$$V(p_t) = V(p_0 + (p_1 - p_0) + (p_2 - p_1) + \dots + (p_t - p_{t-1})) \quad (3)$$

$$= V(p_0 + \Delta p_0 + \Delta p_1 + \dots + \Delta p_{t-1}) \quad (4)$$

$$= V(p_0) + \sum_{i=0}^{t-1} \text{Cov}(p_0, \Delta p_i) + \sum_{i=0}^{t-1} \text{Var}(\Delta p_i) + \sum_{0 \leq i < j}^{t-1} \text{Cov}(\Delta p_i, \Delta p_j). \quad (5)$$

Each allele frequency change is equally like to be positive as it is to be negative; thus by symmetry this second term is zero. Additionally $V(p_0) = 0$, as we treat p_0 as a fixed initial frequency. We can write,

$$V(p_t) = \sum_{i=0}^{t-1} \text{Var}(\Delta p_i) + \sum_{0 \leq i < j}^{t-1} \text{Cov}(\Delta p_i, \Delta p_j). \quad (6)$$

The second term, the cumulative impact of variance in allele frequency change can be partitioned into heritable fitness and drift components (Buffalo and Coop 2019; Santiago and Caballero 1995)

$$V(p_t) = \sum_{i=0}^{t-1} \text{Var}(\Delta_D p_i) + \sum_{i=0}^{t-1} \text{Var}(\Delta_H p_i) + \sum_{0 \leq i < j}^{t-1} \text{Cov}(\Delta p_i, \Delta p_j). \quad (7)$$

where $\Delta_H p_t$ and $\Delta_D p_t$ indicate the allele frequency change due to heritable fitness variation and drift respectively. Then, sum of drift variances in allele frequency change is

$$\sum_{i=0}^{t-1} \text{Var}(\Delta_D p_i) = \sum_{i=0}^{t-1} \frac{p_i(1-p_i)}{2N} \quad (8)$$

replacing the heterozygosity in generation i with its expectation, we have

$$\sum_{i=0}^{t-1} \text{Var}(\Delta_D p_i) = p_0(1-p_0) \sum_{i=0}^{t-1} \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^i \quad (9)$$

$$= p_0(1-p_0) \left[1 - \left(1 - \frac{1}{2N}\right)^t\right] \quad (10)$$

which is the usual variance in allele frequency change due to drift. Then, the total allele frequency change from generations 0 to t is $\text{Var}(\tilde{p}_t - \tilde{p}_0) = \text{Var}(\tilde{p}_t) + \text{Var}(\tilde{p}_0) - 2\text{Cov}(\tilde{p}_t, \tilde{p}_0)$, where the covariance depends on the nature of the sampling plan (see Nei and Tajima 1981; Waples 1989). In the case where there is heritable variation for fitness, and using the fact that $\text{Cov}(\tilde{p}_t, \tilde{p}_0) = p_0(1-p_0)/2N$ for Plan I sampling procedures (Waples 1989), we write,

$$\text{Var}(\tilde{p}_t - \tilde{p}_0) = \text{Var}(\tilde{p}_t) + \text{Var}(\tilde{p}_0) - 2C \text{Cov}(\tilde{p}_t, \tilde{p}_0) \quad (11)$$

$$= \frac{p_t(1-p_t)}{d_t} + \frac{p_0(1-p_0)}{d_0} + p_0(1-p_0) \left[1 - \left(1 - \frac{1}{2N} \right)^t \right] + \quad (12)$$

$$\sum_{i=0}^{t-1} \text{Var}(\Delta_H p_i) + \sum_{0 \leq i < j}^{t-1} \text{Cov}(\Delta p_i, \Delta p_j) - \frac{C p_0(1-p_0)}{2N} \quad (13)$$

$$\frac{\text{Var}(\tilde{p}_t - \tilde{p}_0)}{p_0(1-p_0)} = 1 + \frac{p_t(1-p_t)}{p_0(1-p_0)d_t} + \frac{1}{d_0} - \left(1 - \frac{1}{2N} \right)^t + \quad (14)$$

$$\sum_{i=0}^{t-1} \frac{\text{Var}(\Delta_H p_i)}{p_0(1-p_0)} + \sum_{0 \leq i < j}^{t-1} \frac{\text{Cov}(\Delta p_i, \Delta p_j)}{p_0(1-p_0)} - \frac{C}{N} \quad (15)$$

where $C = 1$ if Plan I is used, and $C = 0$ if Plan II is used (see Waples 1989, p. 380 and Figure 1 for a description of these sampling procedures). We move terms creating a corrected estimator for the population variance in allele frequency change, and replace all population heterozygosity terms with the unbiased sample estimators, e.g. $\frac{d_t}{d_t-1} \tilde{p}_t(1-\tilde{p}_t)$,

$$\frac{d_0-1}{d_0} \frac{\text{Var}(\tilde{p}_1 - \tilde{p}_0)}{\tilde{p}_0(1-\tilde{p}_0)} - \frac{(d_0-1)}{d_0(d_1-1)} \frac{\tilde{p}_1(1-\tilde{p}_1)}{\tilde{p}_0(1-\tilde{p}_0)} - \frac{1}{d_0} + \frac{C}{N} = \frac{\text{Var}(\Delta_H p_0)}{p_0(1-p_0)} + \frac{1}{2N} \quad (16)$$

4.1 Individual and depth sampling process

$X_t \sim \text{Binom}(n_t, p_t)$ where X_t is the count of alleles and n_t is the number of diploids sampled at time t . Then, these individuals are sequenced at a depth of d_t , and $Y_t \sim \text{Binom}(d_t, X_t/n_t)$ reads have the tracked allele. We let $\tilde{p}_t = Y_t/d_t$ be the observed sample allele frequency. Then, the sampling noise is

$$\text{Var}(\tilde{p}_t | p_t) = \mathbb{E}(\text{Var}(\tilde{p}_t | X_t)) + \text{Var}(\mathbb{E}(\tilde{p}_t | X_t)) \quad (17)$$

$$= p_t(1-p_t) \left(\frac{1}{n_t} + \frac{1}{d_t} - \frac{1}{n_t d_t} \right). \quad (18)$$

(see also Jónás et al. 2016).

$$\text{Var}(\tilde{p}_t - \tilde{p}_0) = p_t(1-p_t) \left(\frac{1}{n_t} + \frac{1}{d_t} - \frac{1}{n_t d_t} \right) + p_0(1-p_0) \left(\frac{1}{n_0} + \frac{1}{d_0} - \frac{1}{n_0 d_0} \right) \quad (19)$$

$$- \frac{C p_0(1-p_0)}{N} + p_0(1-p_0) \left[1 - \left(1 - \frac{1}{2N} \right)^t \right] + \sum_{i=0}^{t-1} \text{Var}(\Delta_H p_i) \quad (20)$$

$$+ \sum_{0 \leq i < j}^{t-1} \text{Cov}(\Delta p_i, \Delta p_j) \quad (21)$$

Through the law of total expectation, one can find that an unbiased estimator of the heterozygosity is

$$\frac{n_t d_t}{(n_t - 1)(d_t - 1)} \tilde{p}_t(1 - \tilde{p}_t) \quad (22)$$

$$\begin{aligned} \text{Var}(\tilde{p}_t - \tilde{p}_0) &= \frac{n_t d_t \tilde{p}_t(1 - \tilde{p}_t)}{(n_t - 1)(d_t - 1)} \left(\frac{1}{n_t} + \frac{1}{d_t} - \frac{1}{n_t d_t} \right) + \frac{n_0 d_0 \tilde{p}_0(1 - \tilde{p}_0)}{(n_0 - 1)(d_0 - 1)} \left(\frac{1}{n_0} + \frac{1}{d_0} - \frac{1}{n_0 d_0} \right) + \quad (23) \\ &\quad \frac{n_0 d_0 \tilde{p}_0(1 - \tilde{p}_0)}{(n_0 - 1)(d_0 - 1)} \left[1 - \left(1 - \frac{1}{2N} \right)^t \right] - \frac{C}{N} \frac{n_0 d_0 \tilde{p}_0(1 - \tilde{p}_0)}{(n_0 - 1)(d_0 - 1)} + \\ &\quad \sum_{i=0}^{t-1} \text{Var}(\Delta_H p_i) + \sum_{0 \leq i < j}^{t-1} \text{Cov}(\Delta p_i, \Delta p_j) \\ &= \tilde{p}_t(1 - \tilde{p}_t) \frac{d_t + n_t - 1}{(n_t - 1)(d_t - 1)} + \tilde{p}_0(1 - \tilde{p}_0) \frac{d_0 + n_0 - 1}{(n_0 - 1)(d_0 - 1)} + \quad (24) \\ &\quad \tilde{p}_0(1 - \tilde{p}_0) \frac{n_0 d_0}{(n_0 - 1)(d_0 - 1)} \left[1 - \left(1 - \frac{1}{2N} \right)^t \right] - \frac{C}{N} \tilde{p}_0(1 - \tilde{p}_0) \frac{n_0 d_0}{(n_0 - 1)(d_0 - 1)} \\ &\quad + \sum_{i=0}^{t-1} \text{Var}(\Delta_H p_i) + \sum_{0 \leq i < j}^{t-1} \text{Cov}(\Delta p_i, \Delta p_j) \end{aligned}$$

For $t = 1$,

$$\text{Var}(\tilde{p}_1 - \tilde{p}_0) = \tilde{p}_1(1 - \tilde{p}_1) \frac{d_1 + n_1 - 1}{(n_1 - 1)(d_1 - 1)} + \tilde{p}_0(1 - \tilde{p}_0) \frac{d_0 + n_0 - 1}{(n_0 - 1)(d_0 - 1)} + \quad (25)$$

$$\hat{V}_t = \frac{1}{L} \sum_{l=1}^L (\Delta \tilde{p}_{t,l})^2 \quad (26)$$

TODO

4.2 Covariance Correction

We also need to apply a bias correction to the temporal covariances (and possibly the replicate covariances if the initial sample frequencies are all shared).

The basic issue is that $\text{Cov}(\Delta \tilde{p}_t, \Delta \tilde{p}_{t+1}) = \text{Cov}(\tilde{p}_{t+1} - \tilde{p}_t, \tilde{p}_{t+2} - \tilde{p}_{t+1})$, and thus shares the sampling noise of timepoint $t + 1$. Thus acts to bias the covariance by subtracting off the noise variance term of $\text{Var}(\tilde{p}_{t+1})$, so we add that back in.