August 26, 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, and sequenced every ten generations. Using the seven timepoints and ten replicate populations, we estimated a bias-corrected 60 × 60 temporal-replicate variance-covariance matrix (see XXX for details). Since each replicate population was sequenced every ten generations, the timepoints $t_0 = 0$ generations, $t_1 = 10$ generations, $t_2 = 20$ generations, etc., lead to observed allele frequency changes across ten generation chunks, $\Delta p_{t_0}, \Delta p_{t_1}, \ldots, \Delta p_{t_6}$. Consequently, the ten temporal covariance matrices for each of the ten replicate populations have off-diagonal elements of the form $\mathrm{Cov}(\Delta p_{t_0}, \Delta p_{t_1}) = \mathrm{Cov}(p_{t_1} - p_{t_0}, p_{t_2} - p_{t_1}) = \sum_{i=0}^{10} \sum_{j=10}^{20} \mathrm{Cov}(\Delta p_i, \Delta p_j)$. Each diagonal element has the form $\mathrm{Var}(\Delta p_{t_0}) = \sum_{i=0}^{t_0} \mathrm{Var}(\Delta p_i) + \sum_{i \neq j}^{t_0} \mathrm{Cov}(\Delta p_i, \Delta p_j)$, and is thus a combination of the effects of drift and selection, as both the variance in allele frequency changes and cumulative temporal autocovariances terms increase the variance in allele frequency. With sampling each generation, one could more accurately partition the total variance in allele frequency change (Buffalo and Coop 2019); while we cannot directly estimate the contribution of linked selection to the variance in allele frequency change here, the presence of a positive observed covariance between allele frequency change can only be caused linked selection.

Averaging across replicate populations, we find positive temporal covariances that are statistically significant (p < XXX) consistent with linked selection acting to affect allele frequency changes over very short time periods. We visualize these covariances in Figure (A), which depicts the temporal covariances through time, for each of the five rows covariance matrix. Each row represents the temporal covariance $\mathrm{Cov}(\Delta p_s, \Delta p_t)$, between some initial reference generation s (the row of the matrix), and some later timepoint t (the column of the matrix). For each row, the covariances at first are positive, and then decay towards zero as expected when directional selection affects linked variants' frequency trajectories until ultimately linkage disequilibrium and additive genetic variance for decay (Buffalo and Coop 2019). Note that per replicate, the signal is a bit noisier; see Supplementary Figures XXX.

While the presence of positive temporal covariances is consistent with linked selection affecting allele frequencies over time, this measure not easily interpretable. Alternatively, we can quantify the impact of linked selection through the ratio of total covariance in allele frequency change to the total variance in allele frequency change. Since the total variation in allele frequency change can be decomposed into variance and covariance components, $\mathrm{Var}(p_t - p_0) = \sum_{i \neq j} \mathrm{Cov}(\Delta p_i, \Delta p_j) / \mathrm{Var}(p_t - p_0)$, and the covariances are zero when only drift acts, this is a conservative measure of how much of the variance in allele frequency change is caused by linked selection (Buffalo and Coop 2019). Since timepoints every ten generations were sequenced, we cannot directly estimate the temporal covariances

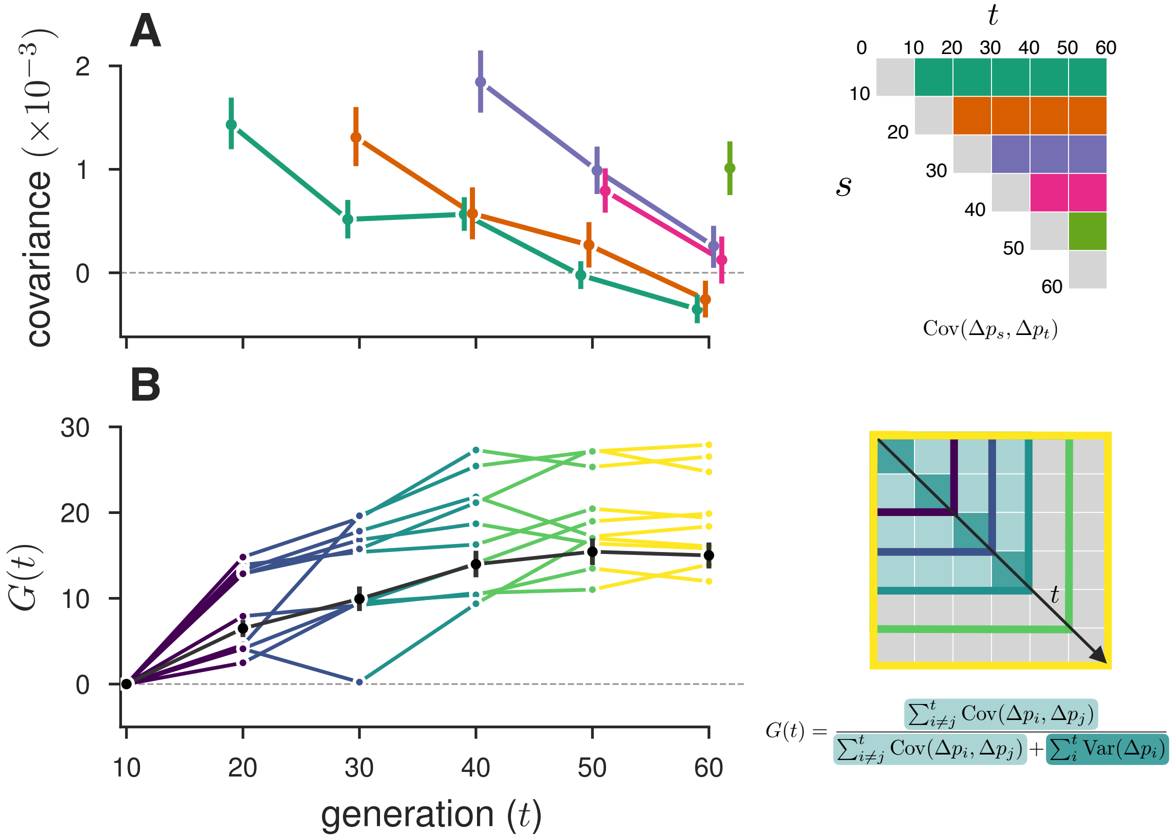

Figure 1: $B = 5000$ bootstraps, with 100

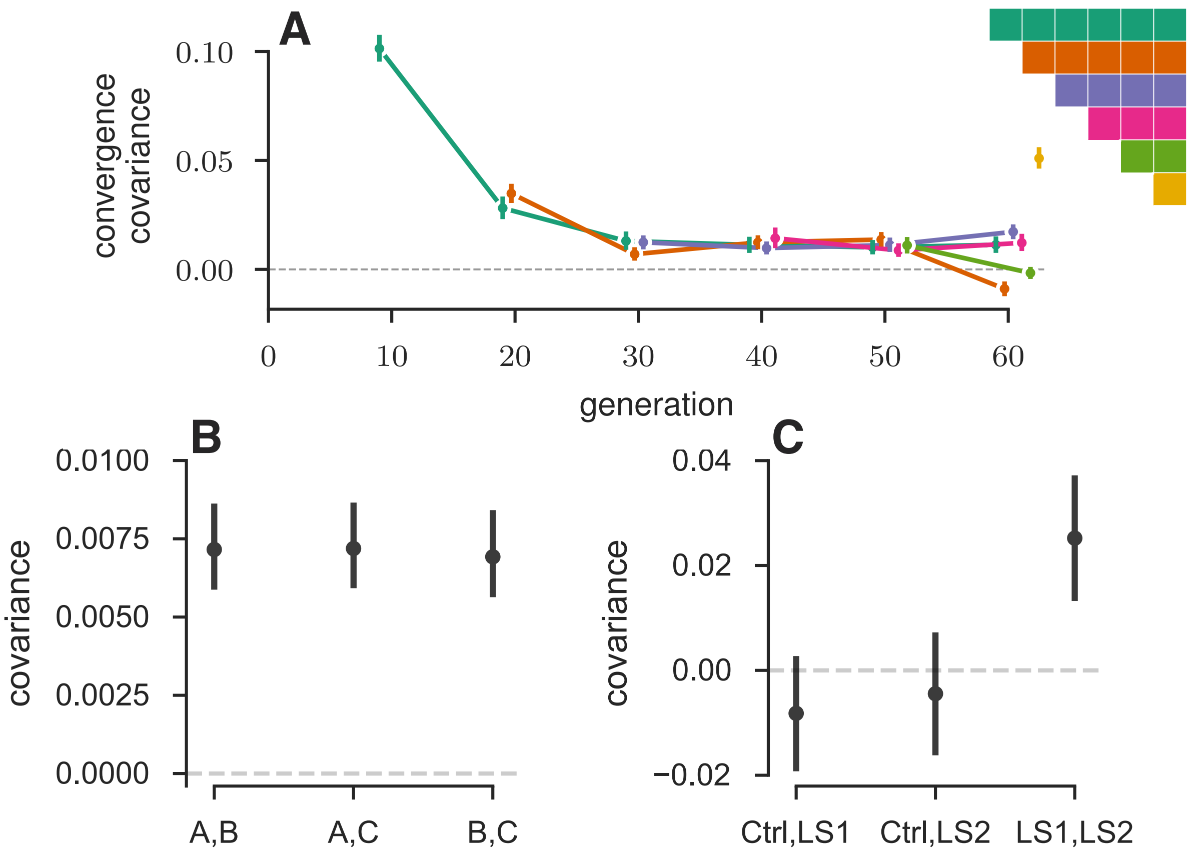

Figure 2

Figure 3

Figure 4

3 Appendix

4 Sampling Bias Corrections

Following Waples (1989), we have that that the variance in the initial generation, which is entirely due to the binomial sampling process, is $\mathrm{Var}(p_0) = {p_0(1-p_0)}/{d_0}$ 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,

$$
\begin{aligned}
\mathrm{Var}(\widetilde{p_t}) &= \mathbb{E}(\mathrm{Var}(\widetilde{p_t}|p_t)) + \mathrm{Var}(\mathbb{E}(\widetilde{p_t}|p_t)) && (1)\\
&= \underbrace{\frac{p_t(1-p_t)}{d_t}}_{\text{generation } t \text{ sampling noise}} + \underbrace{\mathrm{Var}(p_t)}_{\text{variance due to evolutionary process}} && (2)
\end{aligned}
$$

Under a drift-only process, $\mathrm{Var}(p_t) = p_0(1-p_0)\left[1-\left(1-\frac{1}{2N}\right)^t\right]$. 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

$$
\begin{aligned}
V(p_t) &= V\left(p_0 + (p_1 - p_0) + (p_2 - p_1) + \ldots + (p_t - p_{t-1})\right) && (3)\\
&= V\left(p_0 + \Delta p_0 + \Delta p_1 + \ldots + \Delta p_{t-1}\right) && (4)\\
&= V(p_0) + \sum_{i=0}^{t-1} \mathrm{Cov}(p_0, \Delta p_i) + \sum_{i=0}^{t-1} \mathrm{Var}(\Delta p_i) + \sum_{0 \le i < j}^{t-1} \mathrm{Cov}(\Delta p_i, \Delta p_j). && (5)
\end{aligned}
$$

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,

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

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)

$$
V(p_t) = \sum_{i=0}^{t-1} \mathrm{Var}(\Delta_{D} p_i) + \sum_{i=0}^{t-1} \mathrm{Var}(\Delta_{H} p_i) + \sum_{0 \le i < j}^{t-1} \mathrm{Cov}(\Delta p_i, \Delta p_j). \tag{7}
$$

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

$$
\sum_{i=0}^{t-1} \mathrm{Var}(\Delta_{D} p_i) = \sum_{i=0}^{t-1} \frac{p_i(1-p_i)}{2N} \tag{8}
$$

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

$$
\begin{aligned}
\sum_{i=0}^{t-1} \mathrm{Var}(\Delta_{D} p_i) &= p_0(1-p_0) \sum_{i=0}^{t-1} \frac{1}{2N}\left(1 - \frac{1}{2N}\right)^i && (9)\\
&= p_0(1-p_0)\left[1 - \left(1 - \frac{1}{2N}\right)^t\right] && (10)
\end{aligned}
$$

which is the usual variance in allele frequency change due to drift. Then, the total allele frequency change from generations 0 to t is $\mathrm{Var}(\widetilde{p}_t - \widetilde{p}_0) = \mathrm{Var}(\widetilde{p}_t) + \mathrm{Var}(\widetilde{p}_0) - 2\,\mathrm{Cov}(\widetilde{p}_t, \widetilde{p}_0)$, 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 $\mathrm{Cov}(\widetilde{p}_t, \widetilde{p}_0) = {}^{p_0(1-p_0)}/_{2N}$ for Plan I sampling procedures (Waples 1989), we write,

$$\mathrm{Var}(\widetilde{p}_t - \widetilde{p}_0) = \mathrm{Var}(\widetilde{p}_t) + \mathrm{Var}(\widetilde{p}_0) - 2C\,\mathrm{Cov}(\widetilde{p}_t, \widetilde{p}_0) \tag{11}$$

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

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

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

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

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. $\frac{d_t}{d_t - 1}\widetilde{p}_t(1 - \widetilde{p}_t)$,

$$\frac{d_0 - 1}{d_0} \frac{\mathrm{Var}(\widetilde{p}_1 - \widetilde{p}_0)}{\widetilde{p}_0(1 - \widetilde{p}_0)} - \frac{(d_0 - 1)}{d_0(d_1 - 1)} \frac{\widetilde{p}_1(1 - \widetilde{p}_1)}{\widetilde{p}_0(1 - \widetilde{p}_0)} - \frac{1}{d_0} + \frac{C}{N} = \frac{\mathrm{Var}(\Delta_H p_0)}{p_0(1-p_0)} + \frac{1}{2N} \tag{16}$$

4.1 Individual and depth sampling process

$X_t \sim \mathrm{Binom}(n_t, p_t)$ 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 $Y_t \sim \mathrm{Binom}(d_t, {}^{X_t}/_{n_t})$ reads have the tracked allele. We let $\widetilde{p}_t = {}^{Y_t}/_{d_t}$ be the observed sample allele frequency. Then, the sampling noise is

$$\mathrm{Var}(\widetilde{p}_t | p_t) = \mathbb{E}(\mathrm{Var}(\widetilde{p}_t | X_t)) + \mathrm{Var}(\mathbb{E}(\widetilde{p}_t | X_t)) \tag{17}$$

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

(see also Jónás et al. 2016).

$$\begin{aligned}\mathrm{Var}(\widetilde{p}_t - \widetilde{p}_0) = p_t(1-p_t)\left(\frac{1}{n_t} + \frac{1}{d_t} - \frac{1}{n_t d_t}\right) + p_0(1-p_0)\left(\frac{1}{n_0} + \frac{1}{d_0} - \frac{1}{n_0 d_0}\right) && (19)\\ - \frac{Cp_0(1-p_0)}{N} + p_0(1-p_0)\left[1 - \left(1 - \frac{1}{2N}\right)^t\right] + \sum_{i=0}^{t-1} \mathrm{Var}(\Delta_{H} p_i) && (20)\\ + \sum_{0 \le i < j}^{t-1} \mathrm{Cov}(\Delta p_i, \Delta p_j) && (21)\end{aligned}$$

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

$$\frac{n_t d_t}{(n_t - 1)(d_t - 1)} \widetilde{p}_t(1 - \widetilde{p}_t) \tag{22}$$

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

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

4.3 Variance-Covariance Matrix Correction

With frequency collected at $T+1$ timepoints across R replicate populations at L loci, we have $\mathbf{F}$ of allele frequencies, $\mathbf{D}$ multidimensional array of sequencing depths, and a $\mathbf{N}$ multidimensional array of the number of individuals sequenced, each of dimension $R \times (T+1) \times L$. We calculate the array

$\mathbf{\Delta F}$ which contains the allele frequency changes between adjacent generations, and has dimension $R \times T \times L$. The operation flat($\mathbf{\Delta F}$) flattens this array to a $(R \cdot T) \times L$ matrix, such that rows are grouped by replicate, e.g. for timepoint t, replicate r, and locus l and allele frequencies $p_{t,r}$, for a single locus the entries are

$$\text{flat}(\mathbf{\Delta F}) = \begin{bmatrix} \Delta p_{1,0,0} & \Delta p_{2,0,0} & \cdots & \Delta p_{1,1,0} & \Delta p_{2,1,0} & \cdots & \Delta p_{T,R,0} \\ \Delta p_{1,0,1} & \Delta p_{2,0,1} & \cdots & \Delta p_{1,1,1} & \Delta p_{2,1,1} & \cdots & \Delta p_{T,R,1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Delta p_{1,0,L} & \Delta p_{2,0,L} & \cdots & \Delta p_{1,1,L} & \Delta p_{2,1,L} & \cdots & \Delta p_{T,R,L} \end{bmatrix} \tag{25}$$

where each $\Delta p_{t,r,l} = p_{t+1,r,l} - p_{t,r,l}$. Then, the sample temporal-replicate covariance matrix $\mathbf{Q'}$ calculated on flat($\mathbf{\Delta F}$) is a $(R \cdot T) \times (R \cdot T)$ matrix, with the R temporal-covariance block submatrices along the diagonal, and the $R(R-1)$ replicate-covariance submatrices matrices in the upper and lower triangles of the matrix,

$$\mathbf{Q'} = \begin{bmatrix} \mathbf{Q'}_{1,1} & \mathbf{Q'}_{1,2} & \cdots & \mathbf{Q'}_{1,R} \\ \mathbf{Q'}_{2,1} & \mathbf{Q'}_{2,2} & \cdots & \mathbf{Q'}_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Q'}_{R,1} & \mathbf{Q'}_{R,2} & \cdots & \mathbf{Q'}_{R,R} \end{bmatrix} \tag{26}$$

where each submatrix $\mathbf{Q'}_{i,j}$ is the $T \times T$ sample covariance matrix for replicates i and j.

Given the bias of the sample covariance of allele frequency changes, we calculated an expected bias matrix $\mathbf{B}$, averaging over loci,

$$\mathbf{B} = \frac{1}{L} \sum_{l=1}^{L} \frac{\mathbf{h}_l}{2} \circ \left(\frac{1}{\mathbf{d}_l} + \frac{1}{2\mathbf{n}_l} + \frac{1}{2\mathbf{d}_l \circ \mathbf{n}_l} \right) \tag{27}$$

where $\circ$ denotes elementwise product, and $\mathbf{h}_l$, $\mathbf{d}_l$, and $\mathbf{n}_l$, are rows corresponding to locus l of the unbiased heterozygosity arrays $\mathbf{H}$, depth matrix $\mathbf{D}$, and number of diploids matrix $\mathbf{N}$. The unbiased $R \times (T+1) \times L$ heterozygosity array can be calculated as

$$\mathbf{H} = \frac{2\mathbf{D} \circ \mathbf{N}}{(\mathbf{D} - 1) \circ (\mathbf{N} - 1)} \circ \mathbf{F} \circ (1 - \mathbf{F}) \tag{28}$$

where division here is elementwise. Thus, $\mathbf{B}$ is a $R \times (T+1)$ matrix. As explained in XXX, each temporal covariance submatrix $\mathbf{Q}_{r,r}$ requires two corrections.

$$Q_{i,j} = \begin{cases} Q'_{t,s} - b_t - b_s, & \text{if } t = s \\ Q'_{t,s} + b, & \text{if } |i-j| = 1 \end{cases}$$

$Q_{i,j} = \text{Cov}(\Delta p_i, \Delta p_j)$

Additionally, in some study designs, the frequencies from the first timepoint

$$\mathbf{Q} = \mathbf{Q}' - \text{diag}(\text{flat}(\mathbf{B}_{\cdot,2:(T+1)})) - \text{diag}(\text{flat}(\mathbf{B}_{\cdot,1:T})) + \text{offdiag}_1(\text{flat}(\mathbf{B}_{\cdot,2:T})) + \text{offdiag}_{-1}(\text{flat}(\mathbf{B}_{\cdot,2:T})) \tag{30}$$

where diag($\mathbf{x}$) is an operation that takes the vector $\mathbf{x}$ and places them along the diagonal of a matrix. Similarly, offdiag$_k$($\mathbf{x}$) places vector $\mathbf{x}$ along the offdiagonal where for row i, and column j, $j - i = k$. We represent subset of columns s through t of a matrix $\mathbf{B}$ as $\mathbf{B}_{\cdot,s:t}$.

4.4 Block Bootstrap Procedure