Supporting Information for "Skin deep: the decoupling of genetic admixture levels from phenotypes that differed between source populations"

Methods

Quantitative Trait Model: Alternative Approach to Choosing the X_i

In the second case for our quantitative trait model (Edge and Rosenberg 2015a,b), instead of treating allelic type "1" as the trait-increasing "+" allele, we model a trait that is selectively neutral during population divergence: $P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$ for all $i = 1, 2, \ldots, k$, so that at each locus, allelic types "1" and "0" have equal probability of being the "+" allele.

Mating Model: Choosing the Normalizing Coefficients

Recalling that entry m_{ij} in mating probability matrix M gives the probability that a randomly drawn mating pair consists of male i and female j, row and column sums in M must equal $1/N$. For convenience, we work with a doubly stochastic M, dividing it by N just before sampling individuals to obtain a normalized matrix whose entries sum to 1.

We start with an unnormalized mating matrix $\widetilde{M} = [\widetilde{m}_{ij}]$ whose entries are:

$$\widetilde{m}_{ij} = \begin{cases} 1 & \text{random mating} \\ \psi(H_{A,g}^{(i),f}, H_{A,g}^{(j),m}) & \text{assortative mating by admixture} \\ \psi(T_g^{(i),f}, T_g^{(j),m}) & \text{assortative mating by phenotype.} \end{cases} \tag{S1}$$

To construct M from the unnormalized $\widetilde{M}$, we must obtain N^2 normalizing constants α_{ij} so that $M = [m_{ij}] = [\alpha_{ij}\widetilde{m}_{ij}]$ is doubly stochastic. The double stochasticity requirement produces a constraint for each row and column of matrix M. Each entry in M is nonnegative, and m_{ij} must therefore lie in $[0, 1]$.

Infinitely many matrices satisfy the constraints, as the set of $2N$ equations with N^2 variables is underdetermined. Following Ireland and Kullback (1968), we choose M by identifying the matrix that satisfies the set of constraints and that is closest to model matrix $\widetilde{M}$ by the principle of minimum discrimination information (Kullback 1997, pp. 36-43). "Closeness" of two matrices is measured by Kullback-Leibler divergence D_{KL} (Kullback 1997, pp. 1-11), which is nonnegative, equaling zero if and only if the matrices are identical.

The problem of identifying M can be written as a convex optimization problem. We seek to minimize

$$\min_{\{m_{ij}\}} D_{KL}(M||\widetilde{M}) = \min_{\{m_{ij}\}} \sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij} \log \frac{m_{ij}}{\widetilde{m}_{ij}}, \tag{S2}$$

with constraints

$$\begin{aligned} &\sum_{j=1}^{N} m_{ij} = 1 \quad \text{for each } i \text{ from 1 to } N, \\ &\sum_{i=1}^{N} m_{ij} = 1 \quad \text{for each } j \text{ from 1 to } N, \\ &0 \le m_{ij} \le 1 \text{ for all } (i,j) \in \{1, 2, \cdots, N\}^2. \end{aligned} \tag{S3}$$

We use the interior-point method (Nesterov and Nemirovskii 1994; Forsgren 2002), which iteratively traverses the feasible region to obtain the optimal solution numerically, as implemented in the `mosek` function of R package `Rmosek` (MOSEK ApS 2017). For fixed $\widetilde{M}$, the Hessian of the KL divergence has

$$\frac{\partial^2 D_{KL}(M||\widetilde{M})}{\partial m_{ij} \partial m_{k\ell}} = \frac{1}{m_{ij}} \delta_{ik} \delta_{j\ell},$$

where δ is the Kronecker delta. Because $\nabla^2 D_{KL} > 0$ for all $m_{ij} \in (0, 1)$, the KL divergence function is strictly convex (Boyd and Vandenberghe 2004, Section 3.1.4) in each of the N^2 variables in M for fixed $\widetilde{M}$, and numerical minimization finds the unique global minimum (Boyd and Vandenberghe 2004, Section 4.2.1).

The normalization procedure produces relatively little distortion between unnormalized and normalized matrices. In Figure S7, we examine the Kullback-Leibler (KL) divergences and correlation coefficients between unnormalized and normalized matrices in an example setting. The KL divergences stabilize over time at a relatively low value, and the correlation coefficients stabilize near 0.8; a visual illustration of an unnormalized and its associated normalized matrix depicts relatively little difference.

Simulation Procedure: Model for Allele Frequencies

In the second case for our allele frequency model, we simulate sets of k allele frequency pairs $(p_i, q_i), i \in \{1, \ldots, k\}$, following Edge and Rosenberg (2015a). Allele frequencies π_i for derived alleles in the "ancestral" population of S_1 and S_2 are drawn based on the neutral site frequency spectrum: $P[\pi_i = j/(2N_a)] \propto 1/j$, where N_a indicates the size of the ancestral population (Charlesworth and Charlesworth 2010, Eq. B6.6.1). We use $2N_a = 20,000$. We assume each locus i in S_1 and S_2 undergoes independent genetic drift following a split. We add to π_i random numbers $\epsilon_{i,1}$ and $\epsilon_{i,2}$ from a Normal$(0, \gamma\pi_i(1 - \pi_i))$ distribution to simulate derived allele frequencies at locus i in populations S_1 and S_2, respectively. The parameter γ represents the variance introduced by drift into the allele frequencies of the divergent populations. Following Edge and Rosenberg (2015a), we choose $\gamma = 0.3$ so that genetic differentiation between S_1 and S_2 at a group of simulated loci approximates worldwide human F_{ST} estimates. If $\epsilon_{i,1} \geq \epsilon_{i,2}$, then we assign $p_i = \pi_i + \epsilon_{i,1}$ and $q_i = \pi_i + \epsilon_{i,2}$. If $\epsilon_{i,1} < \epsilon_{i,2}$, then we assign $p_i = 1 - (\pi_i + \epsilon_{i,1})$ and $q_i = 1 - (\pi_i + \epsilon_{i,2})$. Note that if this procedure produces $p_i > 1$ or $q_i < 0$, then we assign $p_i = 1$ and $q_i = 0$ so that $0 \leq q_i \leq p_i \leq 1$.

Results

In the main text, we considered the effects of the assortative mating strength c and the number of loci k on the correlation between admixture and phenotype in the admixed population, and on the separate variances of admixture fraction and phenotype. In this Supporting Information, we also examine the effect of the allele frequencies and the manner of assigning trait contributions of individual loci.

Allele Frequencies (p_i and q_i)

We evaluate the effect of allele frequencies p_i and q_i on the quantities of interest. Instead of treating the two source populations as fixed for different alleles, the frequencies p_i and q_i are now sampled according to the simulation procedure described in the "Simulation Procedure: Model for Allele Frequencies" section. Because our results show a monotonic trend across the number of loci we examined ("Number of Trait Loci (k)" section), we focus on a single value of $k = 10$, the number of loci corresponding to the base case.

Cor[H_A, T]

Figure S4A displays Cor[H_A, T] under the model with simulated rather than fixed allele frequencies. Cor[H_A, T] starts from a lower correlation value at time $g = 0$, 0.456, compared to the base case (Figure 4E) value of 1. If all loci have $X_i = 1$, then by definition of trait value T, an individual's trait value is determined by the number of "1" alleles across the trait loci. Because the allele "1" is randomly drawn at each locus $i = 1, 2, \ldots, k$ with probabilities $P(L_{ij} = 1 \mid M = S_1) = p_i$ and $P(L_{ij} = 1 \mid M = S_2) = q_i$ with $j = 1, 2$ ("Quantitative Trait Model" section) and the mean absolute difference between simulated p_i and q_i across k loci is small, some individuals in the source population S_1 have lower trait values than some individuals in S_2, and vice versa. However, due to the constraint $p_i \geq q_i$ across all trait loci, individuals from S_1 have higher probability of having a larger trait value than those from S_2. This property accounts for the nonzero correlation between ancestry and trait present in the source populations outside the base case setting.

The qualitative differences between the three mating models remain similar to the base case, as shown in Figure S4A. All three mating models, however, show an increased rate of decoupling between the admixture fraction and the trait, in that the correlation decreases more rapidly. For random mating, it takes only 4 generations for $\mathrm{Cor}[H_A, T]$ to drop to below half of its starting value, reaching 0.170. The corresponding values under assortative mating by admixture fraction and assortative mating by trait are $g = 5$ ($\mathrm{Cor}[H_A, T] = 0.240$) and $g = 10$ ($\mathrm{Cor}[H_A, T] = 0.220$), respectively. Compared to the base case, the correlation between ancestry and trait in the source population is weaker if the allele frequencies are drawn from the simulation, and thus, the "1" allele does not necessarily trace back to the source population S_1. Under this setting, the decoupling of admixture fraction and phenotype in producing admixed individuals becomes more significant than in the base case.

Var[H_A] and Var[T]

The admixture fraction values at the source populations are not affected by the allele frequencies: $H_A = 1$ and $H_A = 0$ for all individuals in S_1 and S_2, respectively. If $s_{1,0} = s_{2,0} = 0.5$, then $\mathrm{Var}[H_A]$ starts at 0.25 in the founding parental pool, irrespective of the allele frequencies. Comparing Figure S4B and 5E, the $\mathrm{Var}[H_A]$ curves under random mating (red) and assortative mating by admixture (blue) are not affected by the change in allele frequencies p_i and q_i, holding other parameters fixed. Under random mating and assortative mating by admixture, mate choice is independent of the parameters that affect the quantitative trait, and thus, the change in p_i and q_i does not alter the admixture fraction distribution at each generation.

By contrast, under assortative mating by phenotype, $\mathrm{Var}[H_A]$ (green) is affected by the change in the nature of the allele frequencies. $\mathrm{Var}[H_A]$ under assortative mating by phenotype closely follows that under random mating. The simulated allele frequencies have relatively small differences ($\bar{\delta} \approx 0.0509$) between source populations S_1 and S_2. With $X_i = 1$ for all loci, the between-group difference in trait values is small as well, whereas all individuals in S_1 and S_2 still have $H_A = 1$ and $H_A = 0$, respectively. Therefore, with the simulated allele frequencies, the effect on the admixture fraction of assortative mating by phenotype is similar to that in the random mating case. This scenario contrasts with the base case, where allele "1" can be associated with the source population S_1 with certainty, and $\mathrm{Var}[H_A]$ under assortative mating by phenotype behaves similarly to the case of assortative mating by admixture fraction.

With the simulated allele frequencies, $\mathrm{Var}[T] = 0.877$ in the founding parental pool. At $g = 1$, $\mathrm{Var}[T]$ values under random mating and under assortative mating by admixture are 0.784 and 0.825, respectively. Assortative mating by admixture maintains higher $\mathrm{Var}[T]$ than random mating until $g = 8$ and then follows the $\mathrm{Var}[T]$ curve for random mating. By contrast, $\mathrm{Var}[T]$ under assortative mating by trait gradually increases until $g = 13$, at which it achieves its maximum of 0.977, and then decreases to 0.935 at $g = 40$.

Trait Contributions of Individual Loci (X_i)

Returning to the case with fixed allele frequencies of 1 and 0 in the source populations, we next examine the case in which the trait has the property that both alleles have equal probability of being the "+" allele, as described in the "Quantitative Trait Model" section: $P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$ for all $i = 1, 2, \ldots, k$. Figure S5 displays the results using the number of trait loci from the base case, $k = 10$. The qualitative behavior of the result does not depend on the number of loci, with the other parameters fixed.

Cor[H_A, T]

If we let the number of loci with $X_i = 1$ be z, then the number of loci with $X_i = 0$ is $k - z$. Because $p_i = 1$ and $q_i = 0$ across all loci in the base case, by definition of trait value T, the trait value is $2z$ for every individual in S_1 and $2(k - z)$ for every individual in S_2. For a randomly generated set of X_i, $i = 1, 2, \ldots, k$, under $P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$, if $z \neq k - z$, then $\mathrm{Cor}[H_A, T] = 1$ in the founding parental pool H_0^{par}, as shown in Figure S5A. However, compared with the base case, the correlation decays much more rapidly.

With the $P(X_i = 1) \neq 1$ setting, the ancestry and trait are not as tightly coupled in the source populations. However, as in other cases, assortative mating by phenotype preserves the correlation for the longest, and random mating decouples the correlation the fastest of the three mating models.

If the numbers of loci with $X_i = 1$ and $X_i = 0$ are equal ($z = k - z$), then all individuals in the source populations have trait value k irrespective of their origin, and thus, no correlation exists between trait and ancestry in the source population. Hence, $\text{Cor}[H_A, T]$ is 0 in the founding parental pool H_0^{par}, and the correlation remains at 0 throughout the time simulated, irrespective of the mating type (Figure S5D).

Var[H_A] and Var[T]

The panels of Figure S5 display results under $P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$, fixing other parameters as in base case. By the same reasoning as in the "Allele Frequencies (p_i and q_i)" section, the change in parameters involving the quantitative trait does not affect $\text{Var}[H_A]$ under random mating and under assortative mating by admixture fraction. A comparison of Figure S5B and S5E with Figure 5E shows that $\text{Var}[H_A]$ values under assortative mating by admixture fraction are not affected by the change in the X_i.

For $z \neq k - z$, results with $z = 6$ and $k - z = 4$ appear in Figure S5B. Some correlation between the admixture fraction and allele "1" exists in the source populations, and thus, $\text{Var}[H_A]$ for assortative mating by trait (green) somewhat follows that for assortative mating by admixture (blue). However, compared to the base case (Figure 5E), where the "1" allele can be traced back to S_1 with certainty, a more noticeable deviation from the blue curve is observed. As z increases from 6 to 10, the pattern is similar; the quantitative behavior of $\text{Var}[H_A]$ would approach the base case, equivalent to $z = 10$ (green curve in Figure 5E).

In the founding parental pool, $\text{Var}[T] = 4.002$ in our example with $z \neq k - z$. After one generation of mating, $\text{Var}[T]$ drops to 1.997, 2.934, and 2.923, under random mating, assortative mating by admixture fraction, and assortative mating by trait, respectively. From $g = 2$, $\text{Var}[T]$ gradually increases and achieves steady state values for the three models near 4.942 at $g = 5$, 4.934 $g = 8$, and 6.929 at $g = 14$, respectively. In accord with other cases, assortative mating by trait has the highest $\text{Var}[T]$ values across generations.

If $z = k - z$ (Figure S5E), then allele "1" has equal probability of being traced back to either source population. In this scenario, the $\text{Var}[H_A]$ curve from assortative mating follows the $\text{Var}[H_A]$ curve from random mating. Because all individuals have the same trait value in the founding parental pool irrespective of their origin, $\text{Var}[T] = 0$ at $g = 0$. For all three mating models, $\text{Var}[T]$ gradually increases from $g = 1$ to achieve steady state values that are the same as those from the $z \neq k - z$ case.

References

Charlesworth B, Charlesworth D, 2010. *Elements of evolutionary genetics.* Roberts and Company, Greenwood Village, CO.

Edge MD, Rosenberg NA, 2015a. A general model of the relationship between the apportionment of human genetic diversity and the apportionment of human phenotypic diversity. *Human Biology*, 87:313–337.

Edge MD, Rosenberg NA, 2015b. Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 52:32–45.

Forsgren A, Gill PE, Wright MH, 2002. Interior methods for nonlinear optimization. *SIAM Review*, 44:525–597.

Ireland CT, Kullback S, 1968. Contingency tables with given marginals. *Biometrika*, 55:179–188.

Kullback S, 1997. *Information theory and statistics.* Dover Publications, Mineola, NY.

MOSEK ApS, 2017. Mosek rmosek package 8.0.0.78.

Nesterov Y, Nemirovskii A, 1994. *Interior-point polynomial algorithms in convex programming.* Society for Industrial and Applied Mathematics, Philadelphia.

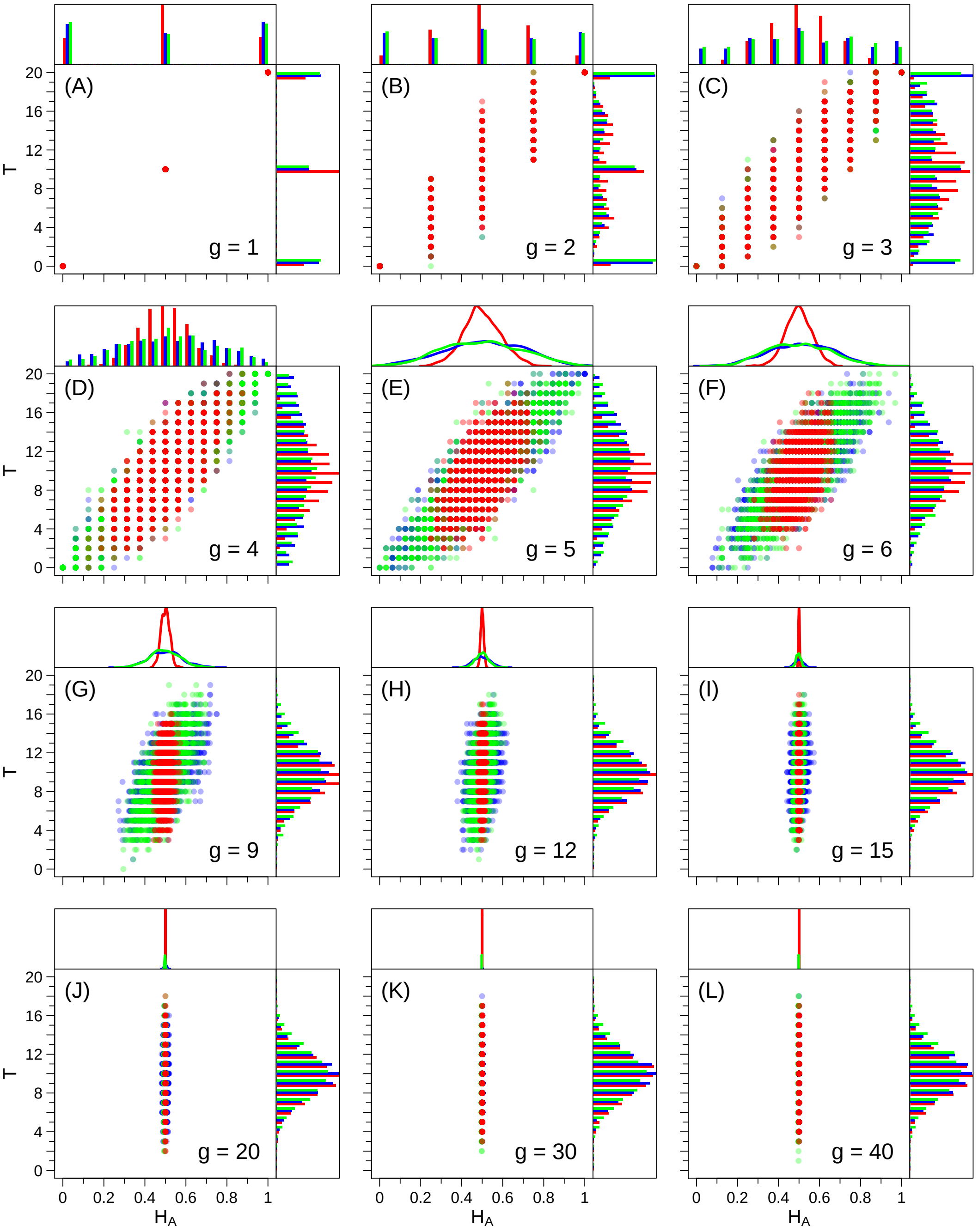

Figure S1: Joint distribution of H_A and T as a function of time. The simulation shown is the same one from Figures 4E, 5E, and 6E, using the parameters from the base case. As described in the "Population Model" and "Quantitative Trait Model" sections, the possible values for the admixture fraction at generation g are $0, 1/2^g, 2/2^g, \ldots, (2^g-1)/2^g, 1$, whereas the possible values for the trait are $0, 1, \ldots, 2k$ across all generations. In each panel, the top, right, and center plots display a marginal distribution of H_A, a marginal distribution of T, and a joint distribution of H_A and T, respectively. Colors follow Figure 4.

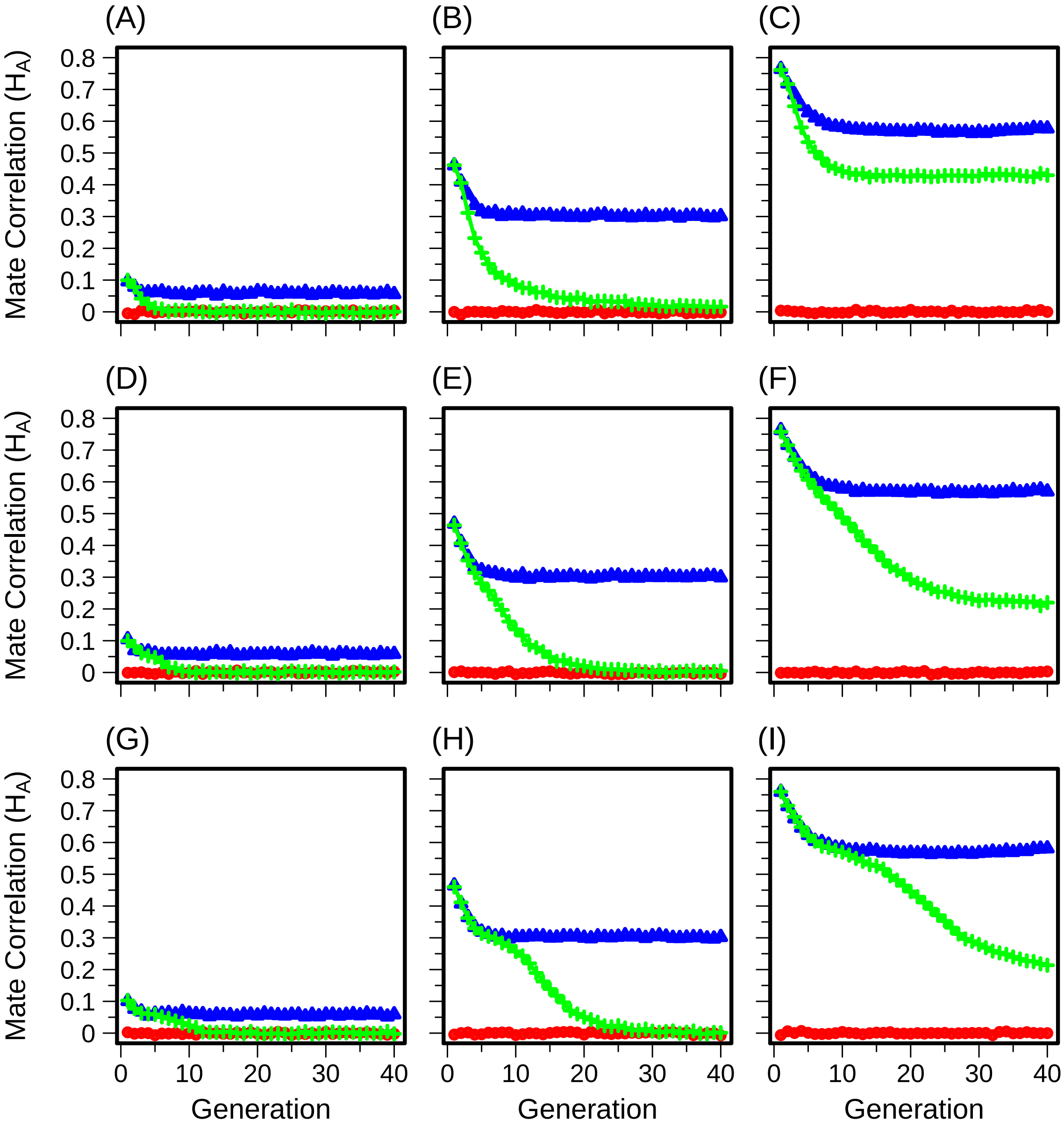

Figure S2: The correlation coefficient $\text{Cor}[H^f_{A,g}, H^m_{A,g}]$ between the admixture fractions of the members of mating pairs as a function of time. The simulations shown are the same ones from Figure 4. (A) $k = 1$, $c = 0.1$. (B) $k = 1$, $c = 0.5$. (C) $k = 1$, $c = 1.0$. (D) $k = 10$, $c = 0.1$. (E) $k = 10$, $c = 0.5$. (F) $k = 10$, $c = 1.0$. (G) $k = 100$, $c = 0.1$. (H) $k = 100$, $c = 0.5$. (I) $k = 100$, $c = 1.0$. Colors and symbols follow Figure 4.

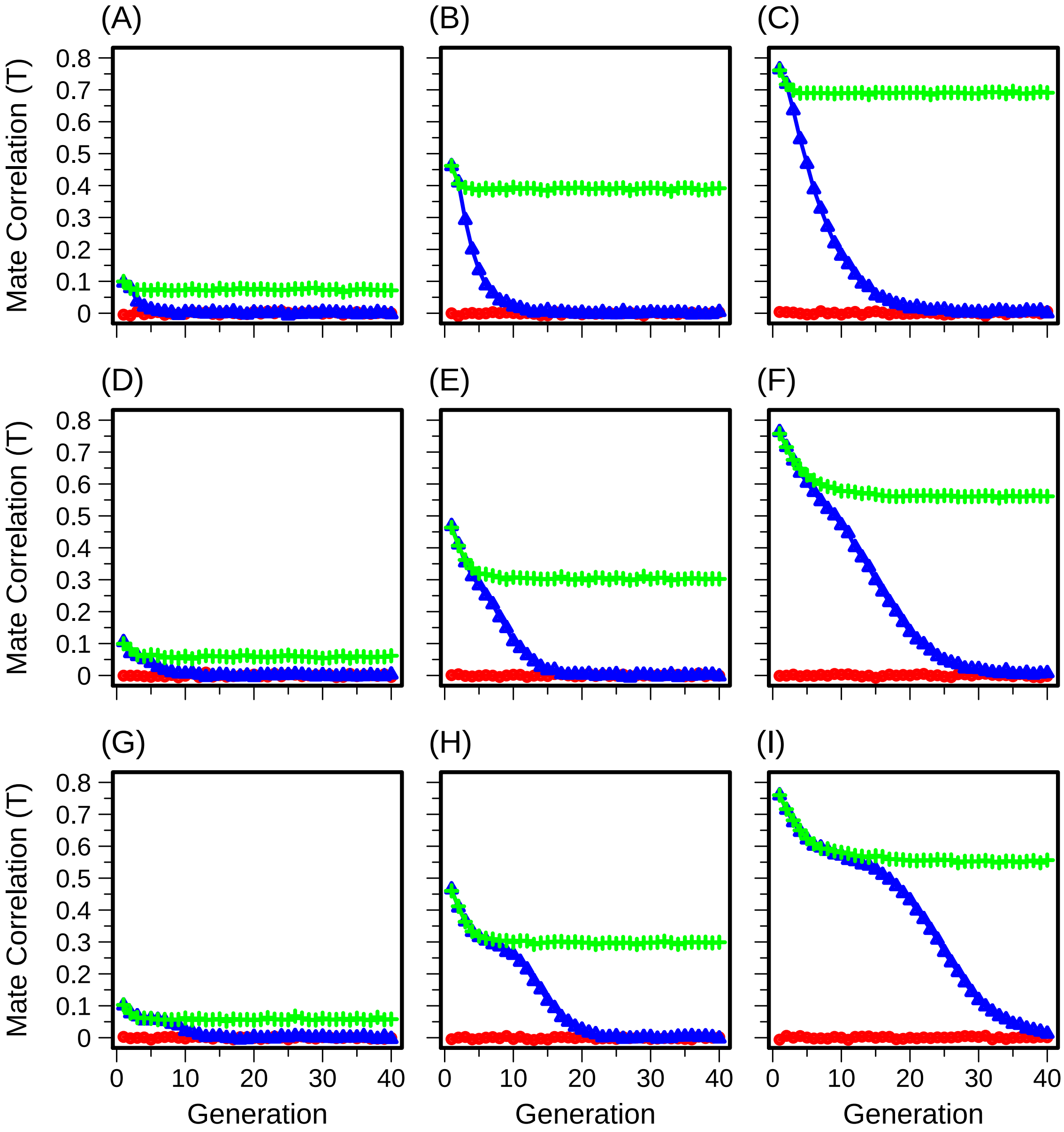

Figure S3: The correlation coefficient $\mathrm{Cor}[T_g^f, T_g^m]$ between the phenotypes of the members of mating pairs as a function of time. The simulations shown are the same ones from Figure 4. (A) $k = 1$, $c = 0.1$. (B) $k = 1$, $c = 0.5$. (C) $k = 1$, $c = 1.0$. (D) $k = 10$, $c = 0.1$. (E) $k = 10$, $c = 0.5$. (F) $k = 10$, $c = 1.0$. (G) $k = 100$, $c = 0.1$. (H) $k = 100$, $c = 0.5$. (I) $k = 100$, $c = 1.0$. Colors and symbols follow Figure 4.

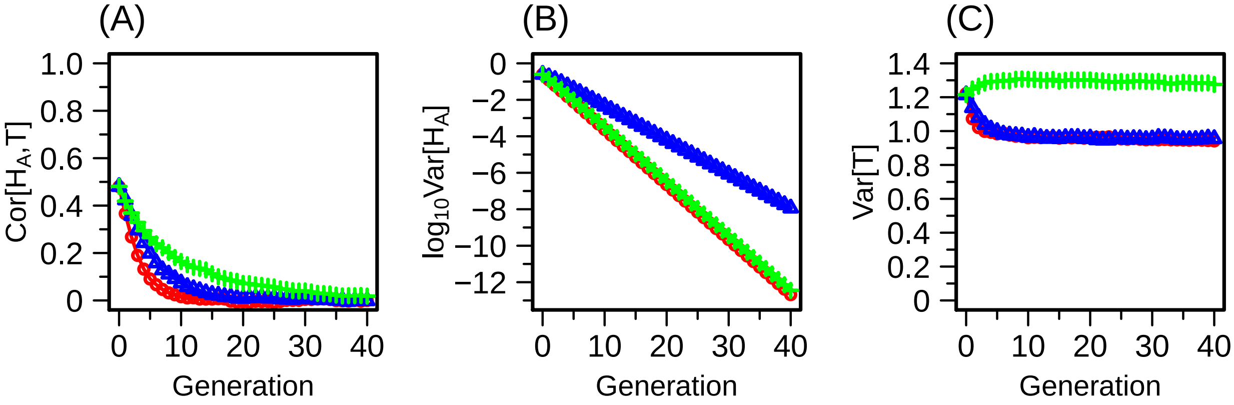

Figure S4: Cor$[H_A, T]$, Var$[H_A]$, and Var$[T]$ in a model in which the allele frequencies p_i and q_i in the source populations S_1 and S_2 are drawn from a simulation rather than being treated as fixed at 1 and 0, respectively. All other parameters are kept at the values of the base case. (A) Correlation between admixture fraction and trait (Cor$[H_A, T]$). (B) Variance of the admixture fraction (Var$[H_A]$). (C) Variance of the phenotype (Var$[T]$). Colors and symbols follow Figure 4. The figure relies on a single replicate of simulated allele frequencies p_i and q_i following a genetic drift model in which S_1 and S_2 descend from a common ancestral population, as described in the "Simulation Procedure: Model for Allele Frequencies" section. The simulated allele frequencies across $k = 10$ loci have mean values of $\bar{p} \approx 0.502$ and $\bar{q} \approx 0.449$ and variance $s_p^2 \approx 0.214$ and $s_q^2 \approx 0.235$. If we let $\delta_i = p_i - q_i$, with $\delta_i > 0$, then the mean of the allele frequency difference across the 10 loci is $\bar{\delta} \approx 5.310 \times 10^{-2}$, with $\overline{\delta^2} \approx 7.865 \times 10^{-3}$. Across $k = 10$ loci, $F_{ST} \approx 0.075$, as computed using Eq. 14 of Edge and Rosenberg (2015a). The y-axis of Var$[H_A]$ is plotted on a logarithmic scale. Results using other replicates of simulated allele frequencies with $k = 10$ are shown in Figure S8.

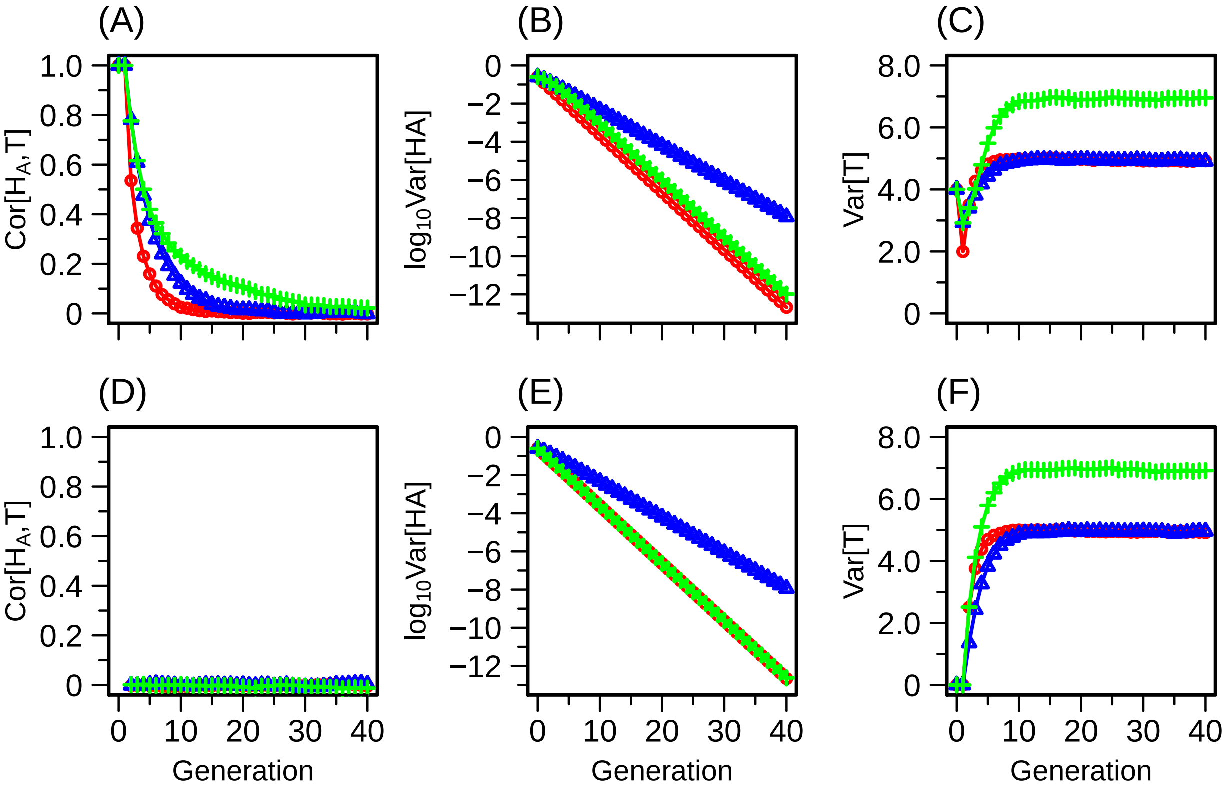

Figure S5: Cor$[H_A, T]$, Var$[H_A]$, and Var$[T]$ under a trait model in which trait loci do not systematically have greater values in one source population: $P(X_i = 1) = P(X_i = 0) = 0.5$. All other parameters are kept at the values of the base case. Of $k = 10$ trait loci, we denote the number of randomly selected loci to have $X_i = 1$ by z. (A) Cor$[H_A, T]$, $z = 6$. (B) Var$[H_A]$, $z = 6$. (C) Var$[T]$, $z = 6$. (D) Cor$[H_A, T]$, $z = 5$. (E) Var$[H_A]$, $z = 5$. (F) Var$[T]$, $z = 5$. Colors and symbols follow Figure 4. Panels A-C and D-F each relies on a single replicate of a set of X_i obtained by sampling the X_i from a Binomial(10, $\frac{1}{2}$) distribution and retaining those with the specified value of $z = 6$ (top panels) and $z = 5$ (bottom panels). The y-axis of Var$[H_A]$ is plotted on a logarithmic scale. Results using other replicates of simulated allele frequencies with $k = 10$ are shown in Figures S9 (for $z = 6$) and S10 (for $z = 5$).

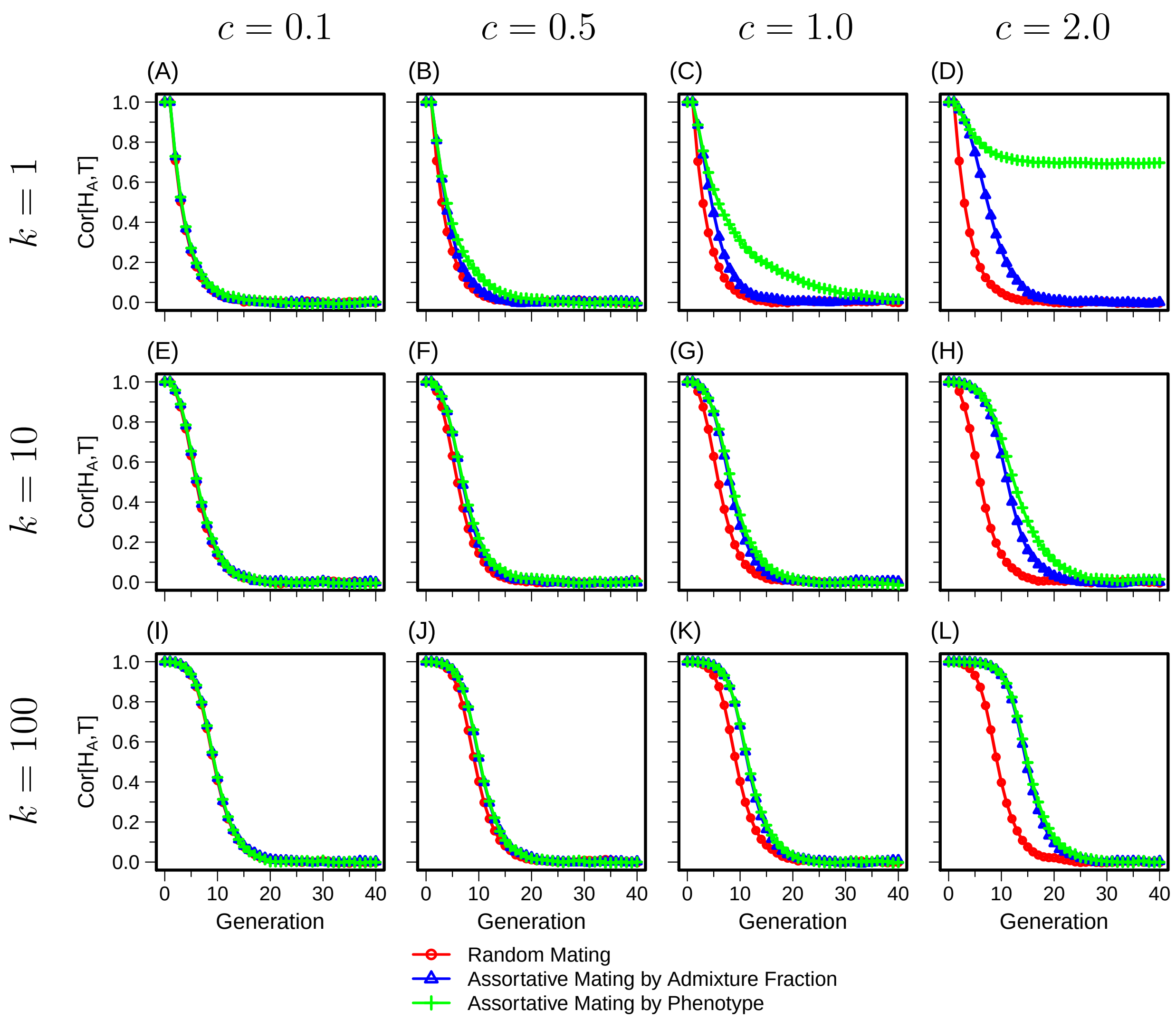

Figure S6: Correlation between admixture fraction and quantitative trait value (Cor$[H_A, T]$) as a function of time, using a fixed rather than a changing scaling factor in the mating function. For the denominator of the exponent in Eq. A2, in assortative mating by admixture, the scaling constant is chosen to be 1 for each generation; for assortative mating by phenotype, the scaling constant is $2k$ for each generation. (A) $k = 1$, $c = 0.1$. (B) $k = 1$, $c = 0.5$. (C) $k = 1$, $c = 1.0$. (D) $k = 1$, $c = 2.0$. (E) $k = 10$, $c = 0.1$. (F) $k = 10$, $c = 0.5$. (G) $k = 10$, $c = 1.0$. (H) $k = 10$, $c = 2.0$. (I) $k = 100$, $c = 0.1$. (J) $k = 100$, $c = 0.5$. (K) $k = 100$, $c = 1.0$. (L) $k = 100$, $c = 2.0$. Colors and symbols follow Figure 4.

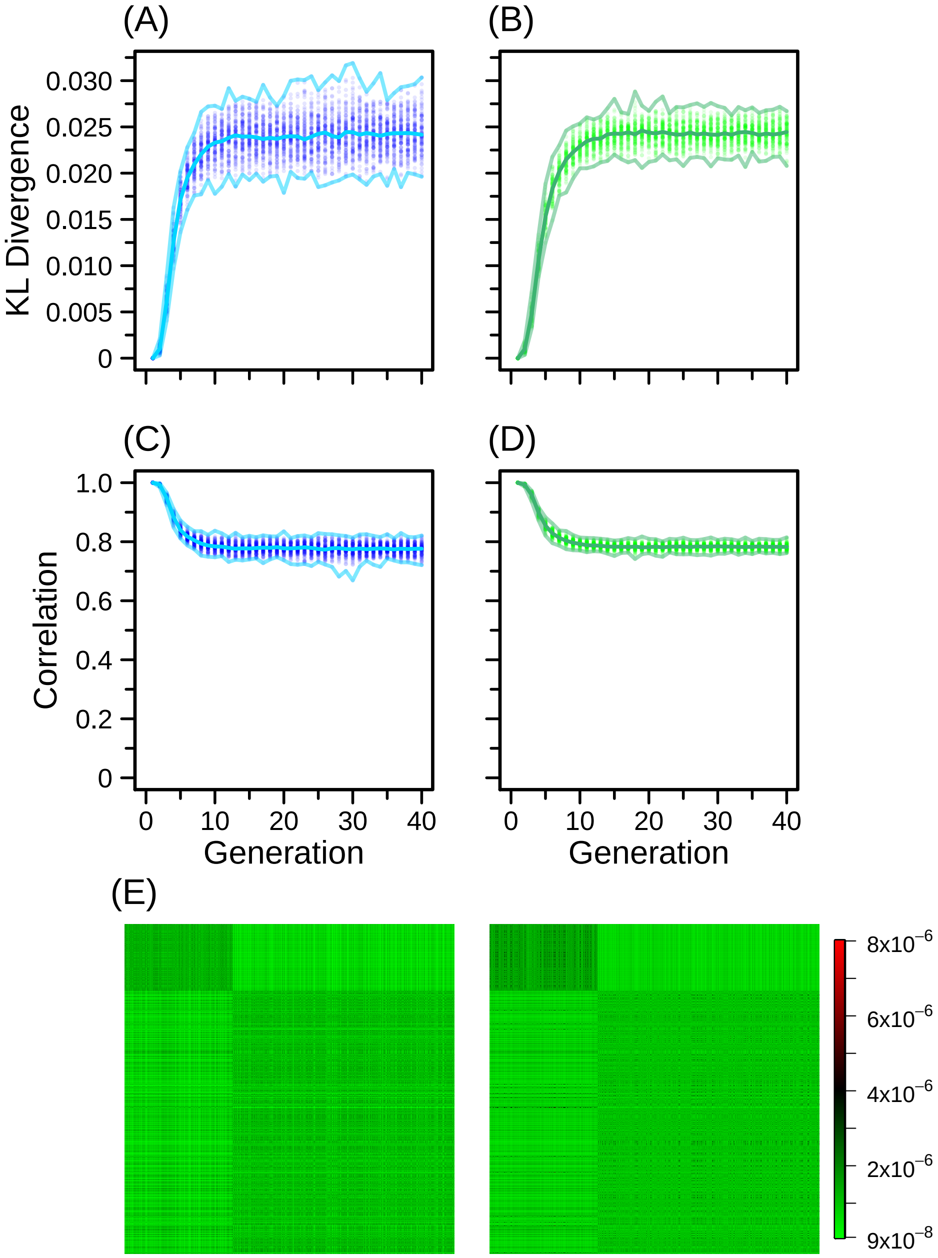

Figure S7: Similarity between unnormalized and normalized mating matrices. The Kullback-Leibler (KL) divergence (Eq. S2) and Pearson correlation coefficient are computed between unnormalized and normalized mating matrices in each generation; the correlation coefficients between matrices are computed by representing a matrix as a vector of its elements. The sum of all entries in normalized matrices M is 1. (A) KL divergence, assortative mating by admixture. (B) KL divergence, assortative mating by phenotype. (C) Pearson correlation coefficient, assortative mating by admixture. (D) Pearson correlation coefficient, assortative mating by phenotype. Each point plotted in a given generation represents a value from one of the 100 replicate trajectories. The lines in each plot represent the minimum, median, and maximum of the 100 replicate values. The simulation shown is the same one from Figures 4E, 5E, and 6E, using the parameters from the base case. The random mating case does not require normalization and is omitted. (E) An example of an unnormalized matrix (left) and the corresponding normalized matrix (right) from an assortative mating by trait simulation replicate with median Pearson correlation coefficient 0.782 at $g = 40$. For ease of visualization, the matrix rows and columns are permuted. The permutation was obtained by biclustering of the unnormalized matrix, and the same permutation was applied to the normalized matrix to preserve row and column orders between two matrices.

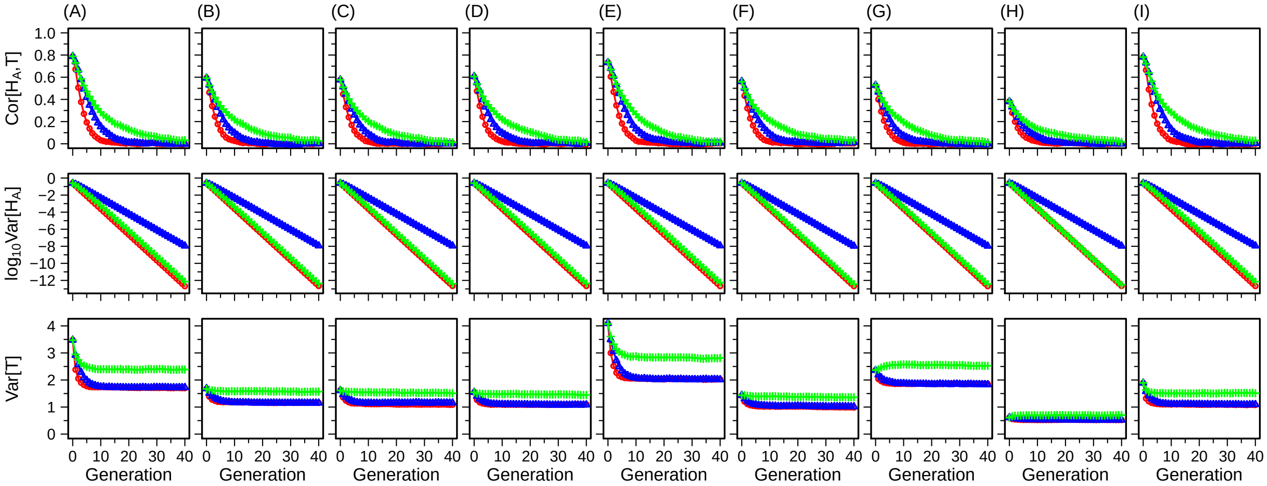

Figure S8: Cor$[H_A, T]$ (top row), Var$[H_A]$ (middle row), and Var$[T]$ (bottom row) using different replicate sets of simulated allele frequencies p_i and q_i with $k = 10$ loci, as described in Figure S4. Different columns represent results from different replicates of simulated p_i and q_i. Colors and symbols follow Figure 4. The y-axis of Var$[H_A]$ is plotted on a logarithmic scale with base 10.

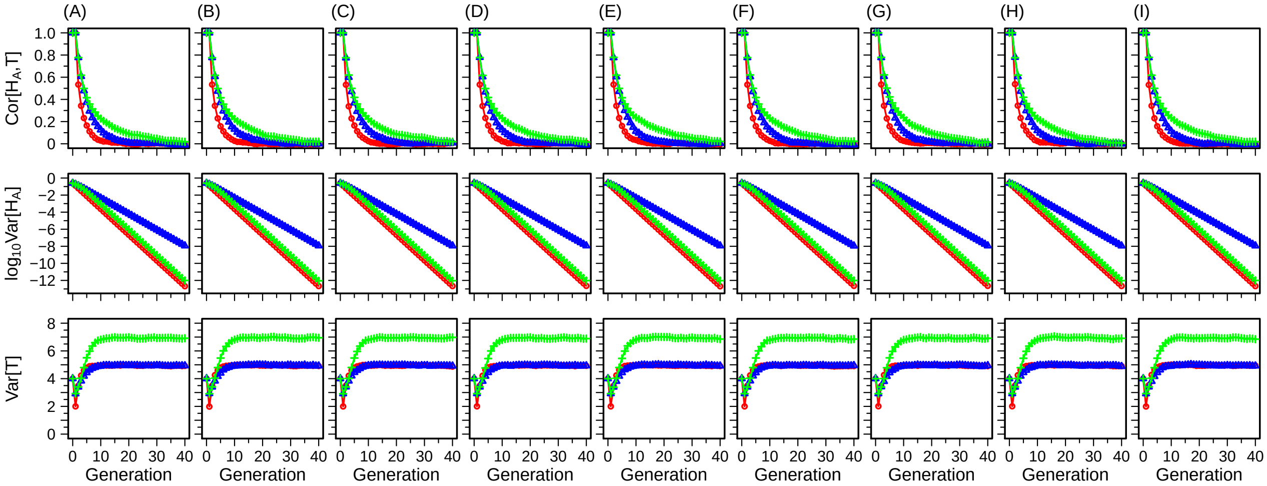

Figure S9: Cor$[H_A, T]$ (top row), Var$[H_A]$ (middle row), and Var$[T]$ (bottom row) using different replicates of a set of X_i for $k = 10$ loci sampled from a Binomial(10, $\frac{1}{2}$) distribution with a constraint $z = 6$ (Figure S5A-C). Colors and symbols follow Figure 4. The y-axis of Var$[H_A]$ is plotted on a logarithmic scale.

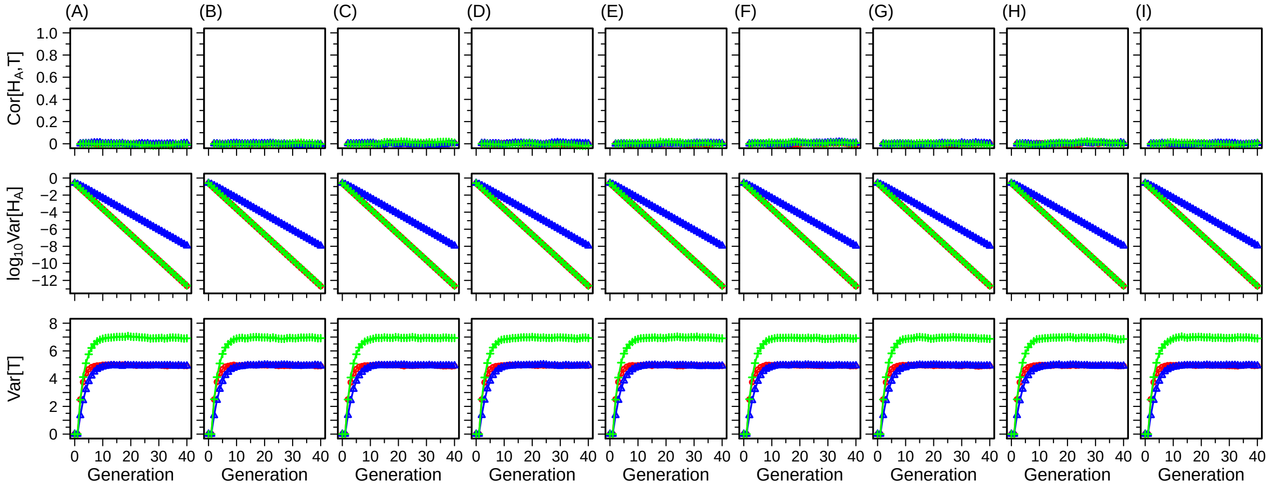

Figure S10: $\mathrm{Cor}[H_A, T]$ (top row), $\mathrm{Var}[H_A]$ (middle row), and $\mathrm{Var}[T]$ (bottom row) using different replicates of a set of X_i for $k = 10$ loci sampled from a Binomial$(10, \frac{1}{2})$ distribution with a constraint $z = 5$ (Figure S5D-F). Colors and symbols follow Figure 4. The y-axis of $\mathrm{Var}[H_A]$ is plotted on a logarithmic scale.