Assortative mating and the dynamical decoupling of genetic admixture levels from phenotypes that differ between source populations

Jaehee Kim[1], Michael D. Edge[2], Amy Goldberg[3], and Noah A. Rosenberg[1]

[1] *Department of Biology, Stanford University, Stanford, CA 94305, USA*
[2] *Department of Evolution and Ecology, University of California, Davis, Davis, CA 95616, USA*
[3] *Department of Evolutionary Anthropology, Duke University, Durham, NC 27708, USA*

Abstract

Source populations for an admixed population can possess distinct patterns of genotype and phenotype at the beginning of the admixture process. Such differences are sometimes taken to serve as markers of ancestry—that is, phenotypes that are initially associated with the ancestral background in one source population are taken to reflect ancestry in that population. Examples exist, however, in which genotypes or phenotypes initially associated with ancestry in one source population have decoupled from overall admixture levels, so that they no longer serve as proxies for genetic ancestry. We develop a mechanistic model for describing the joint dynamics of admixture levels and phenotype distributions in an admixed population. The approach includes a quantitative-genetic model that relates a phenotype to underlying loci that affect its trait value. We consider three forms of mating. First, individuals might assort in a manner that is independent of the overall genetic admixture level. Second, individuals might assort by a quantitative phenotype that is initially correlated with the genetic admixture level. Third, individuals might assort by the genetic admixture level itself. Under the model, we explore the relationship between genetic admixture level and phenotype over time, studying the effect on this relationship of the genetic architecture of the phenotype. We find that the decoupling of genetic ancestry and phenotype can occur surprisingly quickly, especially if the phenotype is driven by a small number of loci. We also find that positive assortative mating attenuates the process of dissociation in relation to a scenario in which mating is random with respect to genetic admixture and with respect to phenotype. The mechanistic framework suggests that in an admixed population, a trait that initially differed between source populations might be a reliable proxy for ancestry for only a short time, especially if the trait is determined by relatively few loci. The results are potentially relevant in admixed human populations, in which phenotypes that have a perceived correlation with ancestry might have social significance as ancestry markers, despite declining correlations with ancestry over time.

Author Summary

Admixed populations are populations that descend from two or more populations that had been separated for a long time at the beginning of the admixture process. The source populations typically possess distinct patterns of genotype and phenotype. Hence, early in the admixture process, phenotypes of admixed individuals can provide information about the extent to which these individuals possess ancestry in a specific source population. To study correlations between admixture levels and phenotypes that differ between source populations, we construct a genetic and phenotypic model of the dynamical process of admixture. Under the model, we show that correlations between admixture levels and these phenotypes dissipate over time—especially if the genetic architecture of the phenotypes involves only a small number of loci, or if mating in the admixed population is random with respect to both the admixture levels and the phenotypes. The result has the implication that a trait that once reflected ancestry in a specific source population might lose this ancestry correlation. As a consequence, in human

populations, after a sufficient length of time, salient phenotypes that can have social meaning as ancestry markers might no longer bear any relationship to genome-wide genetic ancestry.

1 Introduction

Admixed populations descend from two or more source groups that have long been separated and that likely possessed distinct patterns of genotype and phenotype at the beginning of the admixture process. Among individuals in an admixed population in the generations immediately after its founding, admixture levels from any specific source population are highly heterogeneous [1, 2]. For admixed individuals, measurements of specific genotypes and phenotypes that differ in frequency or distribution between source populations can often provide reasonable estimates of individual levels of genetic ancestry in the particular source populations [3, 4]—and for some phenotypes, such measurements might even be commonly regarded by researchers, societies, or admixed individuals themselves as proxies for overall genetic ancestry [5–7].

Examples exist, however, in which genotypes or phenotypes initially associated with ancestry in one source population are decoupled from overall admixture levels, so that they no longer serve as tight proxies for ancestry [5, 6, 8–13]. For example, in human genetics, consider skin pigmentation and eye color, observable traits for which the phenotypic distribution differs substantially between sub-Saharan African and European populations. In the Cape Verdean admixed population, descended from European and West African sources, measurements of skin pigmentation and eye color are correlated with sub-Saharan African genetic ancestry [11]. At the same time, the correlations between phenotype and ancestry are imperfect; many individuals with a high proportion of sub-Saharan African genetic ancestry have skin pigmentation and eye color traits in a range more typical of individuals with higher European genetic ancestry, and vice versa. Similar patterns of incomplete correlation with overall genetic ancestry hold for genotypes that underlie these phenotypes [11].

How does ancestry level become decoupled from genotype and phenotype in an admixed population? Parra et al. [8] proposed one scenario for this decoupling, using an example of assortative mating by a phenotype correlated with ancestry in Brazil. Parra et al. suggested that in Brazil, assortative mating is largely dependent on "color," a phenotypic measure based to a large extent on skin pigmentation. According to this hypothesis, in a population descended from source groups with substantially different skin pigmentation distributions (say, sub-Saharan Africans and Europeans), similarity according to a phenotype correlated with genetic ancestry (say, *color*) increases the probability that a pair is a mating pair. Mating probabilities for pairs of individuals are more closely related to the phenotype than to overall sub-Saharan African or European genetic admixture levels *per se*. Whereas in the early generations of such a process, the phenotype would strongly reflect genetic ancestry, after a sufficient length of time with assortative mating by the phenotype, phenotypic variation would be maintained, but with similar genetic ancestry distributions for individuals with substantially different phenotype (Figure 1). Only at genes associated with the phenotype and their nearby linked genomic regions would genetic ancestry and the phenotype be associated.

Could genetic ancestry in an admixed population become almost entirely decoupled from the phenotypes that differ between its source populations? This scenario is intriguing, as it would eliminate any connection between visible phenotypic markers of genetic ancestry and the genetic ancestry itself; the phenotype of an individual on a trait such as skin pigmentation would reveal little information about the genetic ancestry of molecular characters in an individual—other than for skin pigmentation genes and their closest genomic

neighbors—nor about the total genomic ancestry of the individual.

We develop a mechanistic model describing the joint dynamics of admixture levels and phenotype distributions in an admixed population. The approach includes a quantitative-genetic model that relates a phenotype to underlying loci that affect its trait value. We consider three forms of mating. First, individuals might mate randomly, or assort independently of the overall genetic admixture level. Second, individuals might assort by a phenotype initially correlated with the genetic admixture level, but that is not identical to it. Third, following studies that have detected evidence of assortative mating by genetic admixture levels or phenotypes that are tightly connected to them [14, 15], individuals might assort by the genetic admixture level itself. Under the model, we explore the relationship between genetic admixture level and phenotype over time, studying the effect of the genetic architecture of the phenotype. We find that the decoupling of genetic ancestry and phenotype is not only possible, but it can occur surprisingly quickly, especially if the quantitative phenotype is driven by a small number of loci. Moreover, assortative mating is not required for genotype and phenotype to become decoupled. Indeed, assortative mating attenuates the process compared with a scenario in which mating is random with respect to admixture and with respect to phenotype.

2 Model

2.1 Population Model

Our mechanistic admixture model closely follows the model of Verdu, Goldberg, and Rosenberg [1, 16, 17], building on earlier related models [18–20]. We start with individuals in each of two isolated source populations, S_1 and S_2. At the founding of an admixed population ($g = 0$), a founding parental pool H_0^{par} is formed, containing fraction $s_{1,0}$ from population S_1 and $s_{2,0}$ from population S_2. That is, a random individual in H_0^{par} originates from population S_1 with probability $s_{1,0}$ and from S_2 with probability $s_{2,0}$. This choice requires $s_{1,0} + s_{2,0} = 1$ and $0 \leq s_{1,0}, s_{2,0} \leq 1$. The individuals in the founding parental pool mate according to a mating model (Section 2.3) and produce generation $g = 1$ of admixed offspring (H_1).

In subsequent generations ($g \geq 1$), in forming an admixed population H_{g+1} at generation $g + 1$, three populations contribute to its parental pool H_g^{par}: the source populations (S_1 and S_2) and the admixed population (H_g) of the previous generation, with fractional contributions $s_{1,g}$, $s_{2,g}$, and h_g, respectively. Here, $s_{1,g}$, $s_{2,g}$, and h_g represent probabilities for a random individual in H_g^{par} to originate from populations S_1, S_2, and H_g, with constraints $s_{1,g} + s_{2,g} + h_g = 1$ and $0 \leq s_{1,g}, s_{2,g}, h_g \leq 1$. Offspring resulting from mating among individuals in the parental pool H_g^{par} define the admixed population H_{g+1}. A schematic of the admixture model appears in Figure 2.

The total admixture fraction represents the proportion of the genome of an individual originating from a specific ancestral population, S_1 or S_2. We denote an individual's admixture fraction from source population S_1 at generation g by $H_{A,g}$, with the A indicating consideration of autosomal genetic loci. Given a pair of individuals with admixture fractions $H_{A,g}^{(1)}$ and $H_{A,g}^{(2)}$, the ancestry of their offspring is deterministically set to the mean of the admixture fractions of the parents: $H_{A,g+1} = \frac{1}{2}\left(H_{A,g}^{(1)} + H_{A,g}^{(2)}\right)$. The possible values for the admixture fraction at generation g are $0, 1/2^g, 2/2^g, \ldots, (2^g - 1)/2^g, 1$.

2.2 Quantitative Trait Model

To model a phenotype, we adopt the quantitative trait model of Edge and Rosenberg [21, 22]. We assume each individual is diploid and that k biallelic autosomal loci, each with the same effect size, additively determine the value of a quantitative trait. At each trait locus, we denote the allelic type more prevalent in population S_1 than in population S_2 as allelic type "1", and the other allelic type as "0". The choice is arbitrary in case the allele frequency is the same in the two populations. A diploid individual's genotype at locus i, $i \in \{1, 2, \ldots, k\}$, and allele j, $j \in \{1, 2\}$, is represented by a random indicator variable L_{ij}:

$$L_{ij} = \begin{cases} 1 & \text{if the allele has type "1"} \\ 0 & \text{if the allele has type "0".} \end{cases}$$

Let M be a random variable representing an individual's population membership, considering individuals only from the source populations S_1 and S_2, and define allele frequencies for allelic type "1" at each locus given the population membership:

$$P(L_{ij} = 1 \mid M = S_1) = p_i$$
$$P(L_{ij} = 1 \mid M = S_2) = q_i.$$

Here, j can be either 1 or 2. Because we define allelic type "1" to be more common in population S_1 than in population S_2, $0 \leq q_i \leq p_i \leq 1$.

An individual's trait value is determined by a sum of contributions across loci. At each locus, we denote an allele that increases the trait value by "+" and the other allele by "−". The total quantitative trait value T of an individual given a multilocus genotype is an individual's total number of "+" alleles. Whether the "1" allelic type or "0" type is the "+" allele at locus i is determined by a random variable X_i, following Edge and Rosenberg [21, 22]:

$$X_i = \begin{cases} 1 & \text{if allelic type "1" is "+" allele at locus } i \\ 0 & \text{if allelic type "0" is "+" allele at locus } i. \end{cases}$$

For a given set of values $\{X_1, X_2, \cdots, X_k\}$ for k quantitative trait loci, the total trait value for a diploid individual is equal to the total number of "+" alleles carried by the individuals, or:

$$T \mid_{\{X_1, X_2, \cdots, X_k\}} = \sum_{\{i: X_i = 1\}} \sum_{j=1}^{2} L_{ij} + \sum_{\{i: X_i = 0\}} \sum_{j=1}^{2} (1 - L_{ij}). \tag{1}$$

This quantity takes values in $\{0, 1, \ldots, 2k\}$. An example of the quantitative trait model appears in Figure 3.

We adopt two scenarios for the X_i. We first consider an idealized case in which the number of "1" alleles is perfectly correlated with the trait value: $P(X_i = 1) = 1$ and $P(X_i = 0) = 0$ for all $i = 1, 2, \ldots, k$, so that allelic type "1" is the "+" allele and allelic type "0" is the "−" allele for all loci. Because we define "1" to be the more frequent allelic type in source population S_1, individuals from S_1 are more likely to have a larger trait value than are individuals from S_2. This scenario considers a case in which the phenotype is systematically different between populations 1 and 2, and is depicted in the diagram in Figure 1.

The second case is detailed in Edge and Rosenberg [21, 22], where the trait is selectively neutral during the population divergence and neither allele is preferentially correlated with the trait allele: $P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$ for all $i = 1, 2, \ldots, k$, so that at each locus, allelic type "1" and allelic type "0" have equal probability of being the "+" allele.

2.3 Mating Model

We next describe three different mating models: (1) random mating, in which any two individuals have the same chance of reproducing, independent of phenotype or ancestry; (2) assortative mating by ancestry, in which the probability that two individuals reproduce depends on their ancestries; and (3) assortative mating by phenotype, in which the probability that two individuals reproduce depends on their trait values.

In each generation g, the parental pool H_g^{par} contains $2N$ individuals, N female and N male. The admixture fraction from source population S_1 and trait value of a female individual i are denoted by $H_{A,g}^{(i),f}$ and $T_g^{(i),f}$, respectively. Analogous quantities for a male j are $H_{A,g}^{(j),m}$ and $T_g^{(j),m}$. We construct an $N \times N$ mating probability matrix M, whose entry m_{ij} represents the probability that a female i and a male j mate:

$$\begin{array}{c} \\ (H_{A,g}^{(1),f}, T_g^{(1),f}) \\ (H_{A,g}^{(2),f}, T_g^{(2),f}) \\ \vdots \\ (H_{A,g}^{(N-1),f}, T_g^{(N-1),f}) \\ (H_{A,g}^{(N),f}, T_g^{(N),f}) \end{array} \begin{array}{c} \begin{array}{cccccc} (H_{A,g}^{(1),m}, T_g^{(1),m}) & (H_{A,g}^{(2),m}, T_g^{(2),m}) & \cdots & (H_{A,g}^{(N-1),m}, T_g^{(N-1),m}) & (H_{A,g}^{(N),m}, T_g^{(N),m}) \end{array} \\ \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1(N-1)} & m_{1N} \\ m_{21} & m_{22} & \cdots & m_{2(N-1)} & m_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{(N-1)1} & m_{(N-1)2} & \cdots & m_{(N-1)(N-1)} & m_{(N-1)N} \\ m_{N1} & m_{N2} & \cdots & m_{N(N-1)} & m_{NN} \end{pmatrix} \end{array}. \tag{2}$$

In the absence of selection, every individual in the population must have the same expected number of offspring irrespective of ancestry or phenotype. We assume that the expected number of offspring of an individual is proportional to the expected number of matings of the individual. This quantity is the sum of mating probabilities across all mates available for an individual. Therefore, the equal-offspring requirement translates into an assumption of equal row sums for females and equal column sums for males in the mating matrix in Eq. 2. Note that this assumption of equal numbers of offspring independent of ancestry and phenotype accords with a standard property of assortative mating models that assortative mating on its own does not alter allele frequencies over time [23–28].

The mating probability m_{ij} in Eq. 2 between female i and male j can be expressed as:

$$m_{ij} = \alpha_{ij} \psi(H_{A,g}^{(i),f}, T_g^{(i),f}, H_{A,g}^{(j),m}, T_g^{(j),m}), \tag{3}$$

where ψ is a function that quantifies the dependence of the mating probability m_{ij} on the ancestry and trait values of the individuals in a pair, and α_{ij} is a normalization constant specific to the mating pair (i, j). The constant α_{ij} is included in order to permit the matrix entries to satisfy the constant row and column sum constraints. Without loss of generality, we choose the constant row and column sums to be 1 so that the mating probability matrix M is doubly stochastic. Procedures to evaluate the α_{ij} appear later in the section. The properties of ψ are determined by a mating model.

In random mating, the mating probability is independent of individual ancestry and trait values, so that

$\psi(H_{A,g}^{(i),f}, T_g^{(i),f}, H_{A,g}^{(j),m}, T_g^{(j),m})$ is constant across all i and j, and m_{ij} has the same value for all mating pairs. Therefore, for all pairs (i, j), each taken from $\{1, 2, \ldots, N\}$, $m_{ij} = \alpha_{ij} = \alpha$ for some constant $\alpha \in [0, 1]$.

In assortative mating by ancestry, the mating probability depends only on the ancestries of potential mates and not on the phenotypes: $\psi(H_{A,g}^{(i),f}, T_g^{(i),f}, H_{A,g}^{(j),m}, T_g^{(j),m}) = \psi(H_{A,g}^{(i),f}, H_{A,g}^{(j),m})$ and $m_{ij} = \alpha_{ij}\psi(H_{A,g}^{(i),f}, H_{A,g}^{(j),m})$. For positive assortment, the mating function ψ has higher values if two individuals have similar ancestries and lower values as the ancestries become more different. For example, in complete assortment, ψ is 1 if the two input parameters have the same value and 0 if the values differ. For negative assortment, the behavior of ψ is reversed compared with the positive assortment case: ψ increases as the difference between the ancestries of the two individuals in a mating pair increases.

In assortative mating by phenotype, the mating probability depends only on the trait values of potential mates and not on the ancestries: $\psi(H_{A,g}^{(i),f}, T_g^{(i),f}, H_{A,g}^{(j),m}, T_g^{(j),m}) = \psi(T_g^{(i),f}, T_g^{(j),m})$ and $m_{ij} = \alpha_{ij}\psi(T_g^{(i),f}, T_g^{(j),m})$. The qualitative requirements for the function ψ are the same as with assortative mating by ancestry, but with the trait values of the mating pair as arguments instead of the ancestries.

We adopt the following form for the mating function:

$$\psi(X_g^{(i),f}, X_g^{(j),m}) = e^{\frac{-c|X_g^{(i),f} - X_g^{(j),m}|}{\sigma_{X_g}}}. \tag{4}$$

The finite constant c quantifies the strength of the assortative mating. For a given pair of values $(X_g^{(i),f}, X_g^{(j),m})$, where $X_g = H_{A,g}$ or $X_g = T_g$, a larger c value results in a lower mating probability, which gives stronger positive assortative mating compared with a smaller c value. For a positive value of c, the function takes a value of 1 if two potential mates have the same ancestry level (or phenotype), and ψ decreases exponentially as the difference between the two individuals increases. A negative value of c indicates negative assortative mating, where two individuals with different ancestry (or phenotype) have a higher probability of mating than do two individuals with similar ancestry. We focus on positive assortative mating.

At each generation g, the admixture fraction $H_{A,g}$ takes values in $\{0, 1/2^g, 2/2^g, \ldots, (2^g - 1)/2^g, 1\}$ (Section 2.1), and the phenotype T_g takes values in $\{0, 1, \ldots, 2k\}$ (Section 2.2). To compare statistics from different mating schemes, we consider variables that are standardized by dividing X_g ($H_{A,g}$ or T_g) by its standard deviation σ_{X_g} based on its distribution in H_g^{par} at each generation g. For the unstandardized variables, because T_g takes a higher value than $H_{A,g}$, the effect of assortative mating by phenotype at the same assortative mating strength c is artificially inflated compared to the effect of assortative mating by admixture fraction.

Having specified the mating function ψ, we now formally state the normalization condition for the mating matrix M: the sum across potential mates of the mating probabilities of a random individual in the parental pool must be 1. Recalling that each entry m_{ij} in M represents the probability that individuals i and j mate, the condition requires the row and column sums in the mating matrix to be 1 for each row and column. We start with an unnormalized mating matrix $\widetilde{M} = [\widetilde{m}_{ij}]$ whose entries are:

$$\widetilde{m}_{ij} = \begin{cases} 1 & \text{random mating} \\ \psi(H_{A,g}^{(i),f}, H_{A,g}^{(j),m}) & \text{assortative mating by ancestry} \\ \psi(T_g^{(i),f}, T_g^{(j),m}) & \text{assortative mating by phenotype.} \end{cases} \tag{5}$$

We must obtain N^2 normalizing constants α_{ij} such that the mating matrix $M = [m_{ij}] = [\alpha_{ij}\widetilde{m}_{ij}]$ satisfies the double stochasticity requirement. This requirement gives $2N$ constraints, one for each row and one for each column. Because each entry in the mating matrix represents the probability for two individuals to mate, we also require m_{ij} to be in $[0, 1]$ for all $(i, j) \in \{1, 2, \ldots, N\}^2$.

Infinitely many matrices satisfy the constraints, as the set of $2N$ equations with N^2 variables is underdetermined. We choose the matrix M by identifying the matrix that satisfies the set of constraints and that is closest to our model matrix $\widetilde{M}$ according to the principle of minimum discrimination information (pp. 36-43 in [29]). Here, the "closeness" of a pair of matrices is measured by the Kullback-Leibler divergence D_{KL} (pp. 1-11 in [29]), which is nonnegative and is equal to zero if and only if the two matrices are identical.

The problem of identifying M can be formally written as a convex optimization problem. The objective function that we seek to minimize is

$$\min_{\{m_{ij}\}} D_{KL}(M||\widetilde{M}) = \min_{\{m_{ij}\}} \sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij} \log \frac{m_{ij}}{\widetilde{m}_{ij}}, \tag{6}$$

and we have constraints

$$\begin{aligned} &\sum_{j=1}^{N} m_{ij} = 1 \quad \text{for each } i \text{ from 1 to } N, \\ &\sum_{i=1}^{N} m_{ij} = 1 \quad \text{for each } j \text{ from 1 to } N, \\ &0 \le m_{ij} \le 1 \text{ for all } (i, j) \in \{1, 2, \cdots, N\}^2. \end{aligned} \tag{7}$$

We use the interior-point method [30,31], which iteratively traverses within the feasible region to obtain the optimal solution numerically, as implemented in `mosek` function of R package `Rmosek` [32]. 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 (Section 3.1.4 in [33]) in each of the N^2 variables in M for fixed $\widetilde{M}$, and thus, the optimal solution found by numerical minimization is the unique global minimum (Section 4.2.1 in [33]).

2.4 Expectation and Variance of the Admixture Fraction

To interpret our simulations of admixture dynamics, we will need a series of results concerning the mean and variance of the admixture fraction in the admixed population. In particular, we derive a relationship between the variance of the admixture fraction and the correlation in admixture levels for members of mating pairs.

Let $H_{A,g}$ be a random variable representing the admixture fraction of an individual chosen at random in the admixed population H_g at generation $g \ge 1$. We denote by $(H_{A,g}^{f,p}, H_{A,g}^{m,p})$ the admixture fractions of the members of a mating pair chosen at random from a parental pool H_g^{par} in generation $g \ge 0$. Here, the superscript p denotes that the individual is from the parental pool. The parental pool H_g^{par}, from which the admixed population H_{g+1} at generation $g + 1$ is formed, consists of populations S_1, S_2, and H_g, with

fractional contributions $s_{1,g}$, $s_{2,g}$, and h_g respectively (Section 2.1 and Figure 2). Because we assume each population (S_1, S_2, and H_g) has equally many males and females, the numbers of males and females each remain constant at N, every individual has the same expected number of offspring, and no sex bias by population of origin exists in parental pairings (Section 2.3), $H_{A,g}^{f,p}$ and $H_{A,g}^{m,p}$ are identically distributed.

Let the random variable Y indicate the population membership of a random individual in H_g^{par}. Then

$$H_{A,g}^{f,p}, H_{A,g}^{m,p} = \begin{cases} H_{A,g} & \text{with } P(Y = H_g) = h_g \\ 1 & \text{with } P(Y = S_1) = s_{1,g} \\ 0 & \text{with } P(Y = S_2) = s_{2,g}. \end{cases} \tag{8}$$

For the expectation of admixture in the parental pool, we have

$$\begin{aligned} E[H_{A,g}^{f,p}] = E[H_{A,g}^{m,p}] = E_Y\Big[E[H_{A,g}^{f,p} \mid Y]\Big] &= \sum_{y \in \{S_1, S_2, H_g\}} P(Y = y) E[H_{A,g}^{f,p} \mid Y = y] \\ &= s_{1,g} + h_g E[H_{A,g}] = s_{1,g} + h_g \mu_g. \end{aligned} \tag{9}$$

As a consequence of Eq. 9, we also have

$$E[(H_{A,g}^{f,p})^2] = E[(H_{A,g}^{m,p})^2] = s_{1,g} + h_g E[H_{A,g}^2] = s_{1,g} + h_g \mu_g^2 + h_g \text{Var}[H_{A,g}] \tag{10}$$

$$\text{Var}[H_{A,g}^{f,p}] = \text{Var}[H_{A,g}^{m,p}] = s_{1,g} + h_g \mu_g^2 + h_g \text{Var}[H_{A,g}] - (s_{1,g} + h_g \mu_g)^2. \tag{11}$$

Here, $\mu_g = E[H_{A,g}]$ indicates the expectation of the admixture fraction of a random individual in the admixed population H_g at generation $g \geq 1$.

The ancestry of an offspring individual is deterministically set to the mean of the admixture fractions of the parents. This choice gives:

$$E[H_{A,g+1}] = E\Big[\frac{1}{2}(H_{A,g}^{f,p} + H_{A,g}^{m,p})\Big] = s_{1,g} + h_g E[H_{A,g}] = s_{1,g} + h_g \mu_g. \tag{12}$$

We obtain the recursion for the variance of the admixture fraction over a single generation as follows:

$$\begin{aligned} \text{Var}[H_{A,g+1}] &= E[H_{A,g+1}^2] - (E[H_{A,g+1}])^2 \\ &= \frac{1}{4} E[(H_{A,g}^{f,p} + H_{A,g}^{m,p})(H_{A,g}^{f,p} + H_{A,g}^{m,p})] - (E[H_{A,g+1}])^2 \\ &= \frac{1}{2}\Big(E[(H_{A,g}^{f,p})^2] + E[H_{A,g}^{f,p} H_{A,g}^{m,p}]\Big) - (E[H_{A,g+1}])^2 \\ &= \frac{1}{2}\Big(E[(H_{A,g}^{f,p})^2] + \text{Cor}[H_{A,g}^{f,p}, H_{A,g}^{m,p}] \text{Var}[H_{A,g}^{f,p}] + (E[H_{A,g}^{f,p}])^2\Big) - (E[H_{A,g+1}])^2 \\ &= \frac{1}{2}(1 + r_{H_A,g})\Big[h_g \text{Var}[H_{A,g}] + \mu_g^2 h_g (1 - h_g) - 2\mu_g h_g s_{1,g} + s_{1,g}(1 - s_{1,g})\Big], \end{aligned} \tag{13}$$

where $r_{H_A,g} = \text{Cor}[H_{A,g}^{f,p}, H_{A,g}^{m,p}]$ denotes the correlation of the admixture fractions in a mating pair. The last step is obtained from Eqs. 9–12. As we will see, the time-varying $r_{H_A,g}$ value in general depends on the parameters of the population model, the quantitative trait model, and the mating model.

For a special case of a single admixture event in which source populations S_1 and S_2 do not contribute to the admixed population after its founding ($s_{1,g} = s_{2,g} = 0$ and $h_g = 1$ for all $g \geq 1$), the expectation of

the admixture fraction stays constant in time (Eq. 12), and the variance reduces to a simple formula:

$$\mathrm{Var}[H_{A,g+1}] = \frac{1}{2}(1 + r_{H_A,g})\mathrm{Var}[H_{A,g}]. \tag{14}$$

Under random mating in an infinite population with no ongoing contributions from the source populations, with $r_{H_A,g} = 0$ for all $g \geq 0$, Eq. 14 reduces to the formula $\mathrm{Var}[H_{A,g}] = s_{1,0}(1 - s_{1,0})/2^g$ of Verdu and Rosenberg [1]. Eq. 14 has also been derived by Zaitlen et al. [27] but under assumptions of constant mating correlation ($r_{H_A,g} = r$) across all generations, no migration, and infinite population size.

3 Simulation

3.1 Simulation Procedure

Having specified the populations of interest, the properties of trait values in the populations, and the mating probabilities for pairs of individuals, we now describe how we simulate populations under the model. At the first time step ($g = 0$), $s_{1,0}N$ and $s_{2,0}N$ males are randomly generated from the source populations S_1 and S_2, respectively, with $s_{1,0} + s_{2,0} = 1$. The corresponding numbers of females $s_{1,0}N$ and $s_{2,0}N$ are randomly drawn from source populations, S_1 and S_2, respectively, constituting the founding parental pool H_0^{par} of $2N$ individuals, with N males and N females. All individuals in source population S_1 have an admixture fraction value of 1, and all individuals in source population S_2 have an admixture fraction value of 0, by definition. For each individual in populations S_1 and S_2, genotypes at each of k quantitative trait loci are then randomly generated on the basis of pre-specified allele frequencies p_i and q_i.

We consider two different distributions for the p_i and q_i. First, we assume that the two source populations display fixed differences at all trait loci, so $p_i = 1$ and $q_i = 0$ for all k loci. In this case, every individual in population S_1 has the "1" allele at all trait loci, and every individual in population S_2 has the "0" allele at all trait loci. In subsequent generations, allele "1" can be traced back to population S_1, and allele "0" to S_2 (Figure 1). This choice for the p_i and q_i models a case in which trait-influencing alleles are initially entirely predictive of ancestry and vice versa.

Second, departing from the idealized model, we simulate sets of k allele frequency pairs $(p_i, q_i), i \in \{1, \ldots, k\}$ following Edge and Rosenberg [22]. 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 (Eq. B6.6.1 in [34]). 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 random numbers $\epsilon_{i,1}$ and $\epsilon_{i,2}$ drawn from a Normal$(0, \gamma\pi_i(1 - \pi_i))$ distribution to π_i to simulate derived allele frequencies at locus i in populations S_1 and S_2, respectively. The parameter γ represents the amount of variance introduced by drift into the allele frequencies of the divergent populations. Following Edge and Rosenberg [22], we choose $\gamma = 0.3$ so that the overall degree of 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 $0 \leq q_i \leq p_i \leq 1$.

Using the mating function in Eq. 4, we compute an unnormalized mating matrix for every pair containing

a male and a female from the parental pool:

$$\widetilde{M} = \begin{pmatrix} 1 & e^{-c\Delta_{1,2}} & \cdots & e^{-c\Delta_{1,N-1}} & e^{-c\Delta_{1,N}} \\ e^{-c\Delta_{2,1}} & 1 & \cdots & e^{-c\Delta_{2,N-1}} & e^{-c\Delta_{2,N}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e^{-c\Delta_{N-1,1}} & e^{-c\Delta_{N-1,2}} & \cdots & 1 & e^{-c\Delta_{N-1,N}} \\ e^{-c\Delta_{N,1}} & e^{-c\Delta_{N,2}} & \cdots & e^{-c\Delta_{N,N-1}} & 1 \end{pmatrix}, \tag{15}$$

where $\Delta_{i,j} = \mid H_{A,0}^{(i),f} - H_{A,0}^{(j),m} \mid / \sigma_{H_{A,0}}$ for assortative mating by ancestry and $\Delta_{i,j} = \mid T_0^{(i),f} - T_0^{(j),m} \mid / \sigma_{T_0}$ for assortative mating by trait. For random mating, $c = 0$ and all entries equal 1. The matrix $\widetilde{M}$ is normalized using the procedure of Section 2.3, producing the mating probability matrix M.

Considering all N^2 potential mating pairs, we randomly draw N mating pairs with replacement from the parental pool, weighting mate choices by the mating probabilities in M. Once mates are chosen, each mating pair produces two children, one male and one female, in order to keep the population size of the offspring generation constant at N males and N females. An admixture fraction for an offspring individual is then assigned as the mean of its parental admixture fractions. Assuming no linkage disequilibrium and no mutation, the genotype of the offspring at the quantitative trait loci is then determined by independently selecting at each locus one random allele from one parent and one from the other. The resulting $2N$ offspring form the admixed population H_1 at generation 1.

In subsequent generations $g \geq 1$, we randomly select $s_{1,g}N$, $s_{2,g}N$, and h_gN males and $s_{1,g}N$, $s_{2,g}N$, and h_gN females from S_1, S_2, and H_g, respectively, forming a gth generation parental pool H_g^{par} of $2N$ individuals, consisting of N males and N females. The procedure to generate the offspring population H_{g+1} from H_g^{par} is the same as the procedure for generating H_1 from H_0^{par}.

Throughout the simulation, we keep the population size parameter N constant at 1,000 for computational efficiency in the matrix normalization step. Here, the admixed population size (N) is not necessarily identical to the source population sizes (N_a). For each set of parameters, $(k, p_1, p_2, \ldots, p_k, q_1, q_2, \ldots, q_k, X_1, X_2, \ldots, X_k, c, s_{1,0}, s_{1,g}, s_{2,g})$, we propagated the population to $G = 40$ generations. We generated 100 independent trajectories for each parameter set. For each trajectory, we computed statistics of interest, averaging them over all 100 trajectories in the simulation given the fixed set of parameters.

3.2 Base Case

We start with an idealized base case. First, we specify the parameters involving the population model (Section 2.1). We assume an equal influx from each source population at founding $g = 0$: $s_{1,0} = 0.5$, $s_{2,0} = 1 - s_{1,0} = 0.5$. We also assume no additional contributions from the source populations in the subsequent generations, $s_{1,g} = s_{2,g} = 0$, and $h_g = 1 - s_{1,g} - s_{2,g} = 1$ for all $g \geq 1$.

Next, we choose parameter values for the quantitative trait model (Section 2.2). We consider $k = 10$ trait loci. Across the k loci, all "1" alleles come from source population S_1 and all "0" alleles come from S_2: $p_i = 1$ and $q_i = 0$ for all $i = 1, 2, \ldots, k$. For each locus i contributing to the quantitative trait, we define "1" to be the "+" allele and "0" to be the "−" allele: $X_i = 1$ for all $i = 1, 2, \ldots, k$.

Finally, for the mating model (Section 2.3), we set the assortative mating strength c in Eq. 4 to 0.5.

3.3 Statistics Measured

In each simulated admixed population, in each generation g, we computed the following statistics: correlation between admixture fraction and trait ($\text{Cor}[H_A, T]$), variance of the admixture fraction in the population ($\text{Var}[H_A]$), and variance of the trait value in the population ($\text{Var}[T]$). In the following section, we discuss how these statistics of interest change as we modify the simulation parameters.

4 Results

4.1 Base Case

4.1.1 Correlation between genetic ancestry and phenotype ($\text{Cor}[H_A, T]$)

In the base case, each individual from S_1 has admixture fraction $H_A = 1$ and trait value $T = k$, and each individual from S_2 has admixture fraction $H_A = 0$ and trait value $T = 0$. Therefore, in the founding parental pool, H_0^{par}, admixture fraction and trait value are perfectly correlated: $\text{Cor}[H_A, T] = 1$. In subsequent generations, however, the correlation between the admixture fraction and the trait values starts to decouple, as illustrated in Figure 1. With all parameters involving the population model and the quantitative trait model fixed, the rate of decay in $\text{Cor}[H_A, T]$ depends on the mating model.

A comparison of $\text{Cor}[H_A, T]$ under the three mating models using base-case parameters appears in Figure 4E. Irrespective of the mating model, the founding parental pool has a perfect correlation between ancestry and phenotype. Even if the population starts with perfect correlation between admixture fractions and trait values, however, then random mating rapidly decouples them (red curve). It takes 6 generations of random mating for the correlation to decrease below 0.5 ($\text{Cor}[H_A, T] = 0.490$). After $g = 20$ generations, the correlation becomes 0.137, and it is near zero at $g = 40$ (-0.003).

Compared to random mating, positive assortative mating slows the decoupling of admixture fractions and trait values. Assortative mating by phenotype (green curve in Figure 4E) maintains the correlation longer than assortative mating by admixture fraction (blue curve in Figure 4E). It takes 11 generations under assortative mating by phenotype for the correlation to drop below $\frac{1}{2}$ ($\text{Cor}[H_A, T] = 0.490$), and 10 generations under assortative mating by admixture ($\text{Cor}[H_A, T] = 0.443$). Across the 40 generations we simulated, $\text{Cor}[H_A, T]$ is consistently higher under assortative mating by phenotype than under assortative mating by admixture fraction. The correlation decreases to 0.227 at $g = 20$ and 0.043 at $g = 40$ under assortative mating by phenotype. The corresponding values under assortative mating by admixture are 0.065 at $g = 20$ and 0.009 at $g = 40$, both considerably lower than under assortative mating by phenotype.

The speed of the decoupling between admixture fraction and trait value increases with two factors: Mendelian noise in generating admixed individuals from the admixed population and decreasing contributions from the source populations. Compared to random mating, both assortative mating models have higher probabilities for matings within source populations, and thus, the proportion of individuals produced in the admixed population at $g = 1$ that are genetically admixed is smaller (blue and green lines in the marginal plots for H_A in Figure S1A). Over time, as displayed in Figure S1, random mating pulls individuals away from the source populations, pushing the H_A and T distributions towards the mean values rapidly. On the other hand, both assortative mating models maintain individuals with H_A and T values near the source population values for longer, and thus, they retain higher $\text{Cor}[H_A, T]$ than random mating.

The difference in $\mathrm{Cor}[H_A, T]$ between the two assortative mating models arises from the difference between $\mathrm{Var}[H_A]$ and $\mathrm{Var}[T]$. $\mathrm{Cov}[H_A, T]$ is similar under the two models. Given the similar covariance,

$$\frac{\mathrm{Cor}[H_A, T]_{\mathrm{phen}}}{\mathrm{Cor}[H_A, T]_{\mathrm{gen}}} \approx \sqrt{\frac{\mathrm{Var}[H_A]_{\mathrm{gen}}}{\mathrm{Var}[H_A]_{\mathrm{phen}}} \times \frac{\mathrm{Var}[T]_{\mathrm{gen}}}{\mathrm{Var}[T]_{\mathrm{phen}}}},$$

where the subscripts "gen" and "phen" indicate the property on which mating pairs assort. As we show in the next section, both assortative mating models increase $\mathrm{Var}[H_A]$ and $\mathrm{Var}[T]$ compared to random mating, and the increase in variance is the largest for the property on which mating assorts: $\mathrm{Var}[H_A]_{\mathrm{gen}} > \mathrm{Var}[H_A]_{\mathrm{phen}}$ and $\mathrm{Var}[T]_{\mathrm{phen}} > \mathrm{Var}[T]_{\mathrm{gen}}$. However, we will see that the increase in $\mathrm{Var}[H_A]$ due to assortative mating by admixture fraction exceeds the increase in $\mathrm{Var}[T]$ due to assortative mating by trait:

$$\frac{\mathrm{Var}[H_A]_{\mathrm{gen}}}{\mathrm{Var}[H_A]_{\mathrm{phen}}} > \frac{\mathrm{Var}[T]_{\mathrm{phen}}}{\mathrm{Var}[T]_{\mathrm{gen}}}.$$

This result leads to higher $\mathrm{Cor}[H_A, T]$ under assortative mating by trait compared to that under assortative mating by admixture fraction.

4.1.2 Variance of Ancestry and Phenotype ($\mathrm{Var}[H_A]$ and $\mathrm{Var}[T]$)

Each individual in S_1 has admixture fraction 1, and each individual in S_2 has admixture fraction 0. In the founding parental pool, $\mathrm{Var}[H_A] = 0.250$ for all three mating models. As discussed in Section 2.4, the variance of the admixture fraction can be understood in relation to the correlation coefficient $\mathrm{Cor}[H^f_{A,g}, H^m_{A,g}]$ of the admixture fractions of members of mating pairs. Figure S2 shows this correlation coefficient for the simulations of Figure 4, and Figure S3 shows the analogous correlation $\mathrm{Cor}[T^f_g, T^m_g]$ of trait values.

Figure 5E then shows the variance of the admixture fraction over time under the three mating models, for the same simulations from Figure 4E with the base case parameters. The $\mathrm{Var}[H_A]$ curves in Figure 5E under the three mating models follow Eq. 13, using the time-varying $r_{H_A,g}$ in Figure S2.

Among the three mating models, $\mathrm{Var}[H_A]$ decreases fastest for random mating. After one generation, $\mathrm{Var}[H_A]$ falls in half (0.125), and it continues to decrease monotonically by half. After 40 generations, the value decreases to 2.118×10^{-13}. The distribution of the admixture fraction concentrates around $H_A = \frac{1}{2}$ at each generation. Because the offspring admixture fraction is the mean of those of its parents, without additional influx from the source populations after the founding event, random mating rapidly drives the admixture fraction away from extreme values (0 or 1) toward the mean value of the parental pool ($\frac{1}{2}$).

Under assortative mating by admixture, pairs with similar admixture fraction have higher mating probabilities than under random mating. The fraction of offspring that are admixed is smaller than under random mating, and the admixture fraction distribution remains close to the extreme values (0 or 1) for longer (Figure S1). Hence, $\mathrm{Var}[H_A]$ is larger under assortative mating by admixture fraction (Figure 5E). Without influx from the source populations, $\mathrm{Var}[H_A]$ eventually decreases to zero, but the decrease is slower than for random mating. $\mathrm{Var}[H_A] = 0.184$ after one generation of assortative mating by admixture fraction, and $\mathrm{Var}[H_A] = 1.118 \times 10^{-8}$ after 40 generations. This result can also be seen in Eq. 13. From generation g to $g+1$, $\mathrm{Var}[H_A]_g$ decreases by a factor of $(1 + r_{H_A,g})/2$. With positive assortative mating by admixture ($r_{H_A,g} > 0$), $\mathrm{Var}[H_A]$ in the next generation is increased compared to the case of random mating ($r_{H_A,g} = 0$).

Under assortative mating by phenotype, $\mathrm{Var}[H_A] = 0.183$ after one generation of assortative mating by

phenotype, and $\text{Var}[H_A] = 4.910 \times 10^{-12}$ after 40 generations. For the first few generations ($g < 5$), because $\text{Cor}[H_A, T]$ is high, the correlation between the admixture fraction of mating pairs, and thus $\text{Var}[H_A]$, is similar under the two assortative mating models, as shown in the comparison of the green and blue curves in Figures S2 and 5. However, because the admixture fraction and phenotype decouple by introduction of the Mendelian noise, mating assortatively by phenotype results in lower $r_{H_A,g}$ than mating assortatively by admixture fraction. In accord with Eq. 13, assortative mating by phenotype produces faster decay in $\text{Var}[H_A]$ with its lower $r_{H_A,g}$ at each generation than assortative mating by admixture fraction.

For the variance of the phenotype, using Eq. 1, all individuals in S_1 and S_2 have trait values of 20 and 0, respectively. Therefore, in the founding parental pool, H_0^{par}, noting that S_1 and S_2 each have 1,000 individuals, this variance has the same constant value of $\frac{2{,}000}{1{,}999} \cdot 10^2 = 100.050$ irrespective of the mating model. Figure 6E displays the variance of the phenotype, which decreases most rapidly under random mating. After one generation of random mating, $\text{Var}[T]$ decreases by half (50.025), and it approaches its steady-state value of ≈ 4.957 after 13 generations. Opposite to what was seen for $\text{Var}[H_A]$, however, assortative mating by trait retains $\text{Var}[T]$ higher for longer than assortative mating by admixture fraction. Similar to the case with $\text{Var}[H_A]$ under assortative mating by admixture fraction, assortative mating by phenotype keeps the trait values close to extreme values for longer than the other two mating models.

Having examined the behavior of $\text{Cor}[H_A, T]$, $\text{Var}[H_A]$, and $\text{Var}[T]$ in the base case, we now explore the effect of the assortative mating strength c and the parameters involving the quantitative trait—the number of loci k, allele frequencies p_i and q_i, and trait contribution at each locus X_i—on these quantities.

4.2 Assortative Mating Strength (c)

4.2.1 $\text{Cor}[H_A, T]$

Each row of Figure 4 illustrates the influence of the assortative mating strength c on $\text{Cor}[H_A, T]$ with a fixed number of trait loci k, and each column depicts the effect of the number of loci k on $\text{Cor}[H_A, T]$ with fixed assortative mating strength c. All parameters other than c and k are held constant at the base case values.

With different assortative mating strengths and numbers of trait loci, $c = 0.1, 0.5, 1.0$ and $k = 1, 10, 100$, the qualitative behavior of $\text{Cor}[H_A, T]$ over time remains the same as in the base case. As before, we observe decay in $\text{Cor}[H_A, T]$ under all three mating models, with random mating decoupling ancestry and trait values the most rapidly. $\text{Cor}[H_A, T]$ remains higher for longer under assortative mating by phenotype than under assortative mating by admixture fraction. The rate of decay and the degree to which the patterns differ across the three mating models depend on the assortative mating strength and the number of loci.

If assortative mating is weak ($c = 0.1$ in Figure 4A, D, G), then the effect of assortative mating is small, and thus, $\text{Cor}[H_A, T]$ under assortative mating by admixture and by phenotype closely follows that under random mating. This pattern is seen irrespective of the number of loci. Note that in the limit of $c = 0$, the assortative mating and random mating models are identical because the mating function in Eq. 4 becomes a constant, the same value for all three mating models.

Comparing panels within rows of Figure 4, the results from random mating are identical, as the assortative mating strength does not affect the random mating model. Under both assortative mating models, however, $\text{Cor}[H_A, T]$ increases with the assortative mating strength. In an extreme case of complete assortment ($c \to \infty$), the correlation would stay constant at 1 across all generations: we start with complete correlation

between admixture and phenotype, and $c \to \infty$ implies that only identical individuals can mate, so that the correlation persists unchanged.

The difference among the three models increases with the assortative mating strength given a fixed number of trait loci. The difference is the greatest if $k = 1$ and $c = 1.0$ (Figure 4C). Even after 40 generations, assortative mating by trait retains a high correlation at 0.788, whereas the corresponding values under random mating and assortative mating by admixture are 0.006 and 0.010, respectively.

4.2.2 Var[H_A] and Var[T]

The plots of Var[H_A] in Figure 5 and Var[T] in Figure 6 consider the same simulations that appear for Cor[H_A, T] in Figure 4. As is seen in classical work [26, 35, 36], compared to random mating, assortative mating increases the variance of the property on which assortment takes place. Thus, the variance of the admixture fraction is increased to a greater extent under mating by admixture fraction than under mating by phenotype. Similarly, the variance of the phenotype is increased to a greater extent under mating by phenotype than under mating by admixture fraction. Both types of assortative mating increase both Var[H_A] and Var[T] compared with random mating.

The variance-increasing effect of the assortative mating is visible when comparing panels within each row. For low assortative mating strength ($c = 0.1$), panels A, D, and G in Figures 5 and 6 show that minimal differences in Var[H_A] and Var[T] exist between mating models. As c increases, for a given number of loci, Figures 5 and 6 display increased differences between random and assortative mating, with maximal separation at the largest assortative mating strength simulated, $c = 1$ (panels C, F, I). The random mating model is unaffected by the assortative mating strength c, as was seen with Cor[H_A, T] in Section 4.2.1.

4.3 Number of Trait Loci (k)

4.3.1 Cor[H_A, T]

A comparison of panels within columns of Figure 4 shows that under random mating, with more loci associated with the phenotype, the ancestry-phenotype correlation is higher and stays high for longer. In other words, it takes longer for H_A and T to become decoupled. In particular, under random mating, the correlation between trait and ancestry falls below 0.5 at $g = 3$ if $k = 1$, $g = 6$ if $k = 10$, and $g = 10$ if $k = 100$, independent of the assortative mating strength.

As the number of loci increases, results from the models with assortative mating by phenotype and by admixture become similar. If $(c, k) = (1, 100)$ (Figure 4I), then it takes 24 generations for Cor[H_A, T] values under the two models to differ by more than 0.1. Corresponding times for $(c, k) = (1, 1)$ (Figure 4C) and $(c, k) = (1, 10)$ (Figure 4F) are $g = 6$ and $g = 15$, respectively. Recall that the admixture fraction represents the probability that a random allele at a random autosomal genetic locus originates from source population S_1, assuming infinitely many loci. In the $k \to \infty$ limit, with the whole genome contributing to the trait, the assortative mating models by admixture and by phenotype would behave in exactly the same way.

4.3.2 Var[H_A] and Var[T]

Comparing panels within columns in Figure 5, for a given assortative mating strength, Var[H_A] under assortative mating by admixture follows the same curve irrespective of the number of loci. Because the

mating probability is independent of trait values if mating assortatively by admixture, k has no effect.

As in the base case (Section 4.1.2), both assortative mating models have higher Var$[H_A]$ and Var$[T]$ than random mating. Of the two assortative mating models, assortative mating by admixture fraction has greater Var$[H_A]$ than assortative mating by trait at each generation. For Var$[T]$, assortative mating by trait has greater values than assortative mating by admixture fraction. As was seen with Cor$[H_A, T]$ (Section 4.3.1), for Var$[H_A]$ and Var$[T]$, the difference between random mating and both assortative mating models increases with k, and the difference between the two assortative mating models diminishes as k increases.

4.4 Allele Frequencies (p_i and q_i)

Departing further from the base case, we next evaluate the effect of the 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 Section 3.1. Because our results show a monotonic trend across the number of loci we examined (Section 4.3), we focus this analysis on a single value of $k = 10$, the number of loci corresponding to the base case.

4.4.1 Cor$[H_A, T]$

Figure 7A 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 $X_i = 1$, as shown in Eq. 1, then 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 $j = 1, 2$ (Section 2.2) 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 7A. 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 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 $g = 5$ (Cor$[H_A, T] = 0.240$) and $g = 10$ (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 effect of the Mendelian noise in decoupling of admixture fraction and phenotype in producing admixed individuals becomes more significant than the base case.

4.4.2 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 Var$[H_A]$ starts at 0.25 in the founding parental pool, irrespective of the allele frequencies. Comparing Figure 7B and 5E, the 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 $g = 40$.

4.5 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 Section 2.2: $P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$ for all $i = 1, 2, \ldots, k$. Figure 8 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.

4.5.1 $\mathrm{Cor}[H_A, T]$

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, the trait value is $2z$ for every individual in S_1 and $2(k - z)$ for every individual in S_2 (Eq. 1). 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 8A. However, compared with the base case (Figure 3.2), 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 Section 4.1-4.4, 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, $\mathrm{Cor}[H_A, T]$ 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 8D).

4.5.2 Var[H_A] and Var[T]

The panels of Figure 8 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 Section 4.4.2, the change in parameters involving the quantitative trait does not affect Var[H_A] under random mating and under assortative mating by admixture fraction. A comparison of Figure 8B and 8E with Figure 5E shows that 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 8B. Some correlation between the admixture fraction and allele "1" exists in the source populations, and thus, 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 Var[H_A] would approach the base case, equivalent to $z = 10$ (green curve in Figure 5E).

In the founding parental pool, Var[T] = 4.002 in our example with $z \neq k - z$. After one generation of mating, 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 $g = 2$, 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 Section 4.1-4.4, assortative mating by trait has the highest Var[T] values across generations.

If $z = k - z$ (Figure 8E), then allele "1" has equal probability of traced back to either source population. In this scenario, the Var[H_A] curve from assortative mating follows the Var[H_A] curve from random mating. Because all individuals have the same trait value in the founding parental pool irrespective of their origin, Var[T] = 0 at $g = 0$. For all three mating models, 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.

5 Discussion

In this paper, we have devised a mechanistic admixture model in which an admixed population is formed from contributions of a pair of mutually isolated source populations under assortative mating, either by admixture level or by a quantitative phenotype. The approach includes a quantitative-genetic model that relates a quantitative phenotype to underlying loci affecting its trait value and, ultimately, to mate choices. The admixture level and the quantitative phenotype are studied using a discrete-time recursion that describes the evolution of the admixed population. Under this model, we have examined the correlation between genetic ancestry and phenotype in the admixed population as a function of time under three mating models: random mating, assortative mating by admixture fraction, and assortative mating by phenotype.

Initially, ancestry and phenotype are coupled, as the source populations differ in phenotype. Random mating then decouples the correlation between ancestry and trait faster than is seen in both assortative mating models (Figure 4), and assortative mating by phenotype maintains the correlation to a greater extent than does assortative mating by admixture (Figure 4). Compared with random mating, in a similar manner to classic assortative mating models [26, 35, 36], the assortative mating increases the population variance of the property on which the assortment is based (Figures 5 and 6). In fact, our Eq. 13 multiplies the variance of admixture in a model without assortment [1] by a factor that increases with positive assortative mating.

Increasing the strength of assortative mating magnifies the difference observed among the models in the

speed at which the correlation declines (Figure 4). Generally similar qualitative patterns are observed if the source populations are regarded as having fixed differences (Figure 4) or merely frequency differences in alleles affecting the quantitative trait (Figures 7 and 8).

A key observation is that, as the number of loci underlying the quantitative trait increases, the difference in trajectories between the two assortative mating models decreases. Because assortative mating by admixture fraction affects all loci, whereas assortative mating by trait affects only trait loci and their genomic neighbors, as the trait is determined by an increasing number of loci, the behavior of assortative mating by trait increasingly follows that of assortative mating by ancestry (Figure 4). As the number of loci affecting the phenotype increases, assortative mating by trait increasingly reflects assortative mating by admixture, and the correlation between ancestry and trait value persists for longer (Figure 4).

Our study was motivated partly by a hypothesis of Parra et al. [8] claiming that assortative mating by the *color* phenotype in Brazil could eventually decouple *color* from genetic ancestry, so that largely separate subpopulations with distinct *color* could eventually possess similar levels of African genetic ancestry. We have seen not only that assortative mating by a quantitative trait that differs between source populations can decouple the phenotype from the genetic ancestry, but that random mating can decouple the phenotype from genetic ancestry as well. Moreover, the decoupling is *slower* with assortative mating than it is with random mating. In an admixed population with assortative mating that is heavily influenced by a salient phenotype (such as *color* in the scenario of Parra et al. [8]), mating by other genetically influenced phenotypes is random, or perhaps less strongly assortative. Thus, in an admixed population, we might expect that among all the traits to which genotypes contribute, traits that have little influence on mating behavior will decouple from ancestry most rapidly. Those traits on which assortment does occur, such as *color* in Brazil, will be the slowest to decouple from ancestry—but under our model, they eventually will do so. Thus, in an admixed population, phenotypes that once reflected ancestry in the source populations might no longer be predictive of genetic ancestry after a sufficient length of time has passed.

The focus of our simulations has been on understanding demographic phenomena, but the model is relevant to efforts to investigate determinants of disease traits in admixed populations. For example, in admixture-mapping studies and in studies of health disparities involving admixed populations, correlations of phenotypes and admixture levels are often computed [37–40]. The mechanistic model can potentially provide insights into the way in which these correlations change over time in scenarios in which specific trait architectures are of interest.

We note that we have examined a model with a single admixture event at the founding of the admixed population. Under this idealized model, even if the founding admixed population starts with a perfect correlation between admixture fraction and trait value, the correlation decreases over time and eventually approaches zero in the absence of further influx from source populations. In principle, the framework can account for continuous influx from the source populations. If ancestry–trait correlation exists in source populations, then such influx would be expected to slow the decoupling between admixture and phenotype in the admixed populations under all three mating models, while qualitatively maintaining their relative order in the rate of decoupling.

Although the theoretical framework we have developed can incorporate various genetic architectures and population admixture processes (including disassortative mating), our model has a number of limitations. First, it does not include sex bias during the admixture event, a phenomenon that often occurs in admixture

processes [17, 41, 42]. A recent genetic model in a scenario with continuing contributions from the sources does allow for sex bias with assortative mating, but with no phenotype and with the assortative mating occurring by population membership—in the admixed population or in one of the sources—rather than by the admixture level itself [28]. Next, we have chosen to model admixture in an individual as the mean of parental admixture levels; this approach does not account for stochasticity during genetic transmission. Our quantitative trait model does not incorporate dominance, spatial positioning of trait loci along a genome, epistasis, environmental effects on the phenotype, or genotype-by-environment interaction. The latter pair of limitations might be particularly important in using the results in human data, as numerous studies have shown that assortative mating often operates on sociocultural traits [43–46]. It will be important to extend our theoretical framework to include additional features of the quantitative-genetic model, such as by incorporating dominance, linkage, variable effect sizes across the loci contributing to the quantitative trait, and varying heritability of the phenotype.

Acknowledgments. We acknowledge support from National Institutes of Health grants R01 HG005855 and F32 GM130050 and National Science Foundation grant BCS-1515127.

References

[1] Verdu P, Rosenberg NA. A general mechanistic model for admixture histories of hybrid populations. Genetics. 2011;189(4):1413–1426. doi:10.1534/genetics.111.132787.

[2] Gravel S. Population genetics models of local ancestry. Genetics. 2012;191(2):607–619. doi:10.1534/genetics.112.139808.

[3] Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, et al. Ethnic-affiliation estimation by use of population-specific DNA markers. The American Journal of Human Genetics. 1997;60(4):957–964.

[4] Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, et al. Estimating African American admixture proportions by use of population-specific alleles. The American Journal of Human Genetics. 1998;63(6):1839–1851. doi:10.1086/302148.

[5] Parra EJ, Kittles RA, Shriver MD. Implications of correlations between skin color and genetic ancestry for biomedical research. Nature Genetics. 2004;36:S54–S60. doi:10.1038/ng1440.

[6] Ruiz-Linares A, Adhikari K, Acuña-Alonzo V, Quinto-Sanchez M, Jaramillo C, Arias W, et al. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. PLOS Genetics. 2014;10(9):1–13. doi:10.1371/journal.pgen.1004572.

[7] Algee-Hewitt BFB. Population inference from contemporary American craniometrics. American Journal of Physical Anthropology. 2016;160(4):604–624. doi:10.1002/ajpa.22959.

[8] Parra FC, Amado RC, Lambertucci JR, Rocha J, Antunes CM, Pena SDJ. Color and genomic ancestry in Brazilians. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(1):177–182. doi:10.1073/pnas.0126614100.

[9] Pimenta JR, Zuccherato LW, Debes AA, Maselli L, Soares RP, Moura-Neto RS, et al. Color and genomic ancestry in Brazilians: a study with forensic microsatellites. Human Heredity. 2006;62(4):190–195. doi:10.1159/000096872.

[10] Leite TKM, Fonseca RMC, de França NM, Parra EJ, Pereira RW. Genomic ancestry, self-reported "color" and quantitative measures of skin pigmentation in Brazilian admixed siblings. PLOS ONE. 2011;6(11):e27162–e27162. doi:10.1371/journal.pone.0027162.

[11] Beleza S, Johnson NA, Candille SI, Absher DM, Coram MA, Lopes J, et al. Genetic architecture of skin and eye color in an African-European admixed population. PLOS Genetics. 2013;9(3):1–15. doi:10.1371/journal.pgen.1003372.

[12] Durso DF, Bydlowski SP, Hutz MH, Suarez-Kurtz G, Magalhães TR, Junho Pena SD. Association of genetic variants with self-assessed color categories in Brazilians. PLOS ONE. 2014;9(1):1–8. doi:10.1371/journal.pone.0083926.

[13] Magalhães da Silva T, Sandhya Rani MR, de Oliveira Costa GN, Figueiredo MA, Melo PS, Nascimento JaF, et al. The correlation between ancestry and color in two cities of Northeast Brazil with contrasting ethnic compositions. European Journal Of Human Genetics. 2014;23:984–989. doi:10.1038/ejhg.2014.215.

[14] Risch N, Choudhry S, Via M, Basu A, Sebro R, Eng C, et al. Ancestry-related assortative mating in Latino populations. Genome Biology. 2009;10(11):1–16. doi:10.1186/gb-2009-10-11-r132.

[15] Zou JY, Park DS, Burchard EG, Torgerson DG, Pino-Yanes M, Song YS, et al. Genetic and socioeconomic study of mate choice in Latinos reveals novel assortment patterns. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(44):13621–13626. doi:10.1073/pnas.1501741112.

[16] Goldberg A, Verdu P, Rosenberg NA. Autosomal admixture levels are informative about sex bias in admixed populations. Genetics. 2014;198(3):1209–1229. doi:10.1534/genetics.114.166793.

[17] Goldberg A, Rosenberg NA. Beyond 2/3 and 1/3: the complex signatures of sex-biased admixture on the X Chromosome. Genetics. 2015;201(1):263–279. doi:10.1534/genetics.115.178509.

[18] Long JC. The genetic structure of admixed populations. Genetics. 1991;127(2):417–428.

[19] Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. The American Journal of Human Genetics. 1995;57(2):455–464.

[20] Guo W, Fung WK, Shi N, Guo J. On the formula for admixture linkage disequilibrium. Human Heredity. 2005;60(3):177–180. doi:10.1159/000090119.

[21] Edge MD, Rosenberg NA. 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. 2015;52:32–45. doi:10.1016/j.shpsc.2014.12.005.

[22] Edge MD, Rosenberg NA. A general model of the relationship between the apportionment of human genetic diversity and the apportionment of human phenotypic diversity. Human Biology. 2015;87(4):313–337. doi:10.13110/humanbiology.87.4.0313.

[23] Jennings HS. The numerical results of diverse systems of breeding. Genetics. 1916;1(1):53–89.

[24] Fisher RA. XV. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh. 1919;52(2):399–433. doi:10.1017/S0080456800012163.

[25] Wright S. Systems of mating. III. Assortative mating based on somatic resemblance. Genetics. 1921;6(2):144–161.

[26] Crow JF, Kimura M. An introduction to population genetics theory. Blackburn Press; 2009.

[27] Zaitlen N, Huntsman S, Hu D, Spear M, Eng C, Oh SS, et al. The effects of migration and assortative mating on admixture linkage disequilibrium. Genetics. 2017;205(1):375–383. doi:10.1534/genetics.116.192138.

[28] Goldberg A, Rastogi A, Rosenberg NA. Assortative mating by population of origin in a mechanistic model of admixture. Theoretical Population Biology. 2019;.

[29] Kullback S. Information theory and statistics. Mineola, NY: Dover Publications; 1997.

[30] Nesterov Y, Nemirovskii A. Interior-point polynomial algorithms in convex programming. Society for Industrial and Applied Mathematics; 1994.

[31] Forsgren A, Gill PE, Wright MH. Interior methods for nonlinear optimization. SIAM Review. 2002;44(4):525–597. doi:10.1137/S0036144502414942.

[32] MOSEK ApS. MOSEK Rmosek Package 8.0.0.78. 2017;.

[33] Boyd S, Vandenberghe L. Convex optimization. Cambridge, UK: Cambridge University Press; 2004.

[34] Charlesworth B, Charlesworth D. Elements of evolutionary genetics. Greenwood Village, CO: Roberts and Co. Publishers; 2010.

[35] Crow JF, Felsenstein J. The effect of assortative mating on the genetic composition of a population. Eugenics Quarterly. 1968;15(2):85–97. doi:10.1080/19485565.1968.9987760.

[36] Felsenstein J. Continuous-genotype models and assortative mating. Theoretical Population Biology. 1981;19(3):341–357. doi:10.1016/0040-5809(81)90025-3.

[37] Tang H, Jorgenson E, Gadde M, Kardia SLR, Rao DC, Zhu X, et al. Racial admixture and its impact on BMI and blood pressure in African and Mexican Americans. Human Genetics. 2006;119(6):624–633. doi:10.1007/s00439-006-0175-4.

[38] Peralta CA, Ziv E, Katz R, Reiner A, Burchard EG, Fried L, et al. African ancestry, socioeconomic status, and kidney function in elderly African Americans: a genetic admixture analysis. Journal of the American Society of Nephrology. 2006;17(12):3491–3496. doi:10.1681/ASN.2006050493.

[39] Gravlee CC, Non AL, Mulligan CJ. Genetic ancestry, social classification, and racial inequalities in blood pressure in Southeastern Puerto Rico. PLOS ONE. 2009;4(9):e6821–e6821. doi:10.1371/journal.pone.0006821.

[40] Non AL, Gravlee CC, Mulligan CJ. Education, genetic ancestry, and blood pressure in African Americans and Whites. American Journal of Public Health. 2012;102(8):1559–1565. doi:10.2105/AJPH.2011.300448.

[41] Wilkins JF. Unraveling male and female histories from human genetic data. Current Opinion in Genetics & Development. 2006;16(6):611–617. doi:10.1016/j.gde.2006.10.004.

[42] Adhikari K, Chacón-Duque JC, Mendoza-Revilla J, Fuentes-Guajardo M, Ruiz-Linares A. The genetic diversity of the Americas. Annual Review of Genomics and Human Genetics. 2017;18(1):277–296. doi:10.1146/annurev-genom-083115-022331.

[43] Kalmijn M. Intermarriage and homogamy: causes, patterns, trends. Annual Review of Sociology. 1998;24(1):395–421. doi:10.1146/annurev.soc.24.1.395.

[44] Watson D, Klohnen EC, Casillas A, Nus Simms E, Haig J, Berry DS. Match makers and deal breakers: analyses of assortative mating in newlywed couples. Journal of Personality. 2004;72(5):1029–1068. doi:10.1111/j.0022-3506.2004.00289.x.

[45] Rosenfeld MJ. Racial, educational and religious endogamy in the United States: a comparative historical perspective. Social Forces. 2008;87(1):1–31. doi:10.1353/sof.0.0077.

[46] Schwartz CR. Trends and variation in assortative mating: causes and consequences. Annual Review of Sociology. 2013;39(1):451–470. doi:10.1146/annurev-soc-071312-145544.

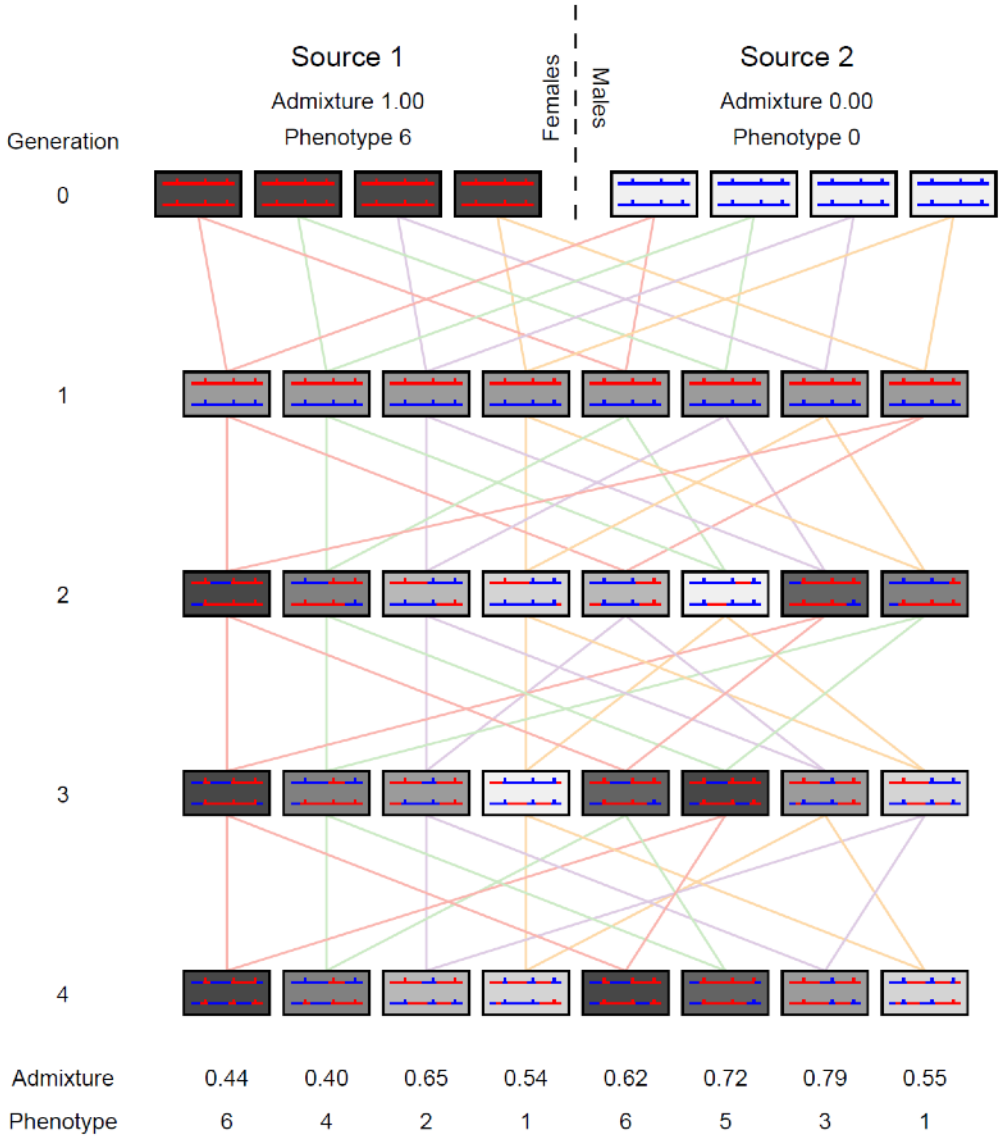

Figure 1: A schematic of an admixture process with positive assortative mating by a phenotype initially correlated with admixture levels. In generation 0, an admixture process begins with females from one population (source 1, left) and males from another (source 2, right). For a quantitative phenotype, source population 1 begins with a high trait value of 6 and source population 2 has a low trait value of 0. Three loci contribute additively to the genetic architecture of the phenotype; each allele derived from source population 1 contributes a value of 1 to the phenotype. The phenotype is represented by the shading of a box. Individuals are depicted as pairs of chromosomes with the ancestral sources of those chromosomes; short vertical lines along the chromosome indicate the three loci that contribute to the phenotype. After generation 1, positive assortative mating by phenotype proceeds in the admixed population. Lines connecting generations are displayed in four colors, representing four mating pairs. Initially, in generation 2, a strong correlation exists between admixture and phenotype ($r = 0.96$). By generation 4, however, owing to recombination events that stochastically dissociate the trait loci from the overall genetic admixture, the genetic admixture has been decoupled from the phenotype, so that some of the individuals with the highest trait values have among the lowest admixture coefficients for source population 1, and the correlation between phenotype and overall genetic admixture has dissipated ($r = -0.09$).

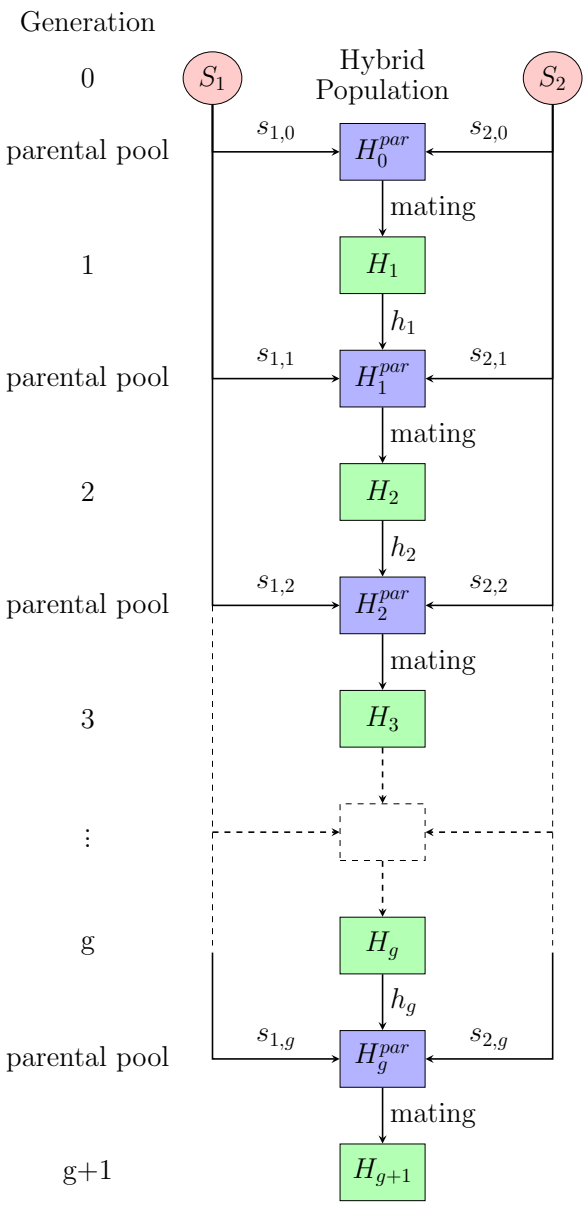

Figure 2: A schematic diagram of the admixture process. At the founding of the population ($g = 0$), two isolated source populations produce the first generation of a admixed population (H_1). In the subsequent generations ($g \geq 1$), populations from S_1, S_2, and H_g constitute a parental pool H_g^{par} at generation g from which the admixed population H_{g+1} at generation $g + 1$ is produced. Fractional contributions from three populations in forming the parental pool are $s_{1,g}$, $s_{2,g}$, and h_g, respectively. Individuals in the parental pool mate based on mating models described in Section 2.3.

	Locus							
	1	2	3	4	5	6	7	8
$\boldsymbol{L_{i1}}$	1	0	0	1	1	1	0	1
$\boldsymbol{L_{i2}}$	0	0	0	1	1	1	1	1
$\boldsymbol{X_i}$	1	1	0	0	0	1	1	0
Contribution to $\boldsymbol{T}$	1	0	2	0	0	2	1	0
	$\boldsymbol{T = 6}$							

Figure 3: An example of our quantitative trait model. Here, a diploid individual with $k = 8$ trait loci is shown. At each locus i, an allele L_{ij} contributes to the overall trait value if and only if $L_{ij} = X_i$ where X_i is a variable indicating which of two alleles, "0" or "1" increases the trait value. The total trait value of an individual equals the number of alleles satisfying $L_{ij} = X_i$ across the k trait loci. In this example, the individual has $T = 6$.

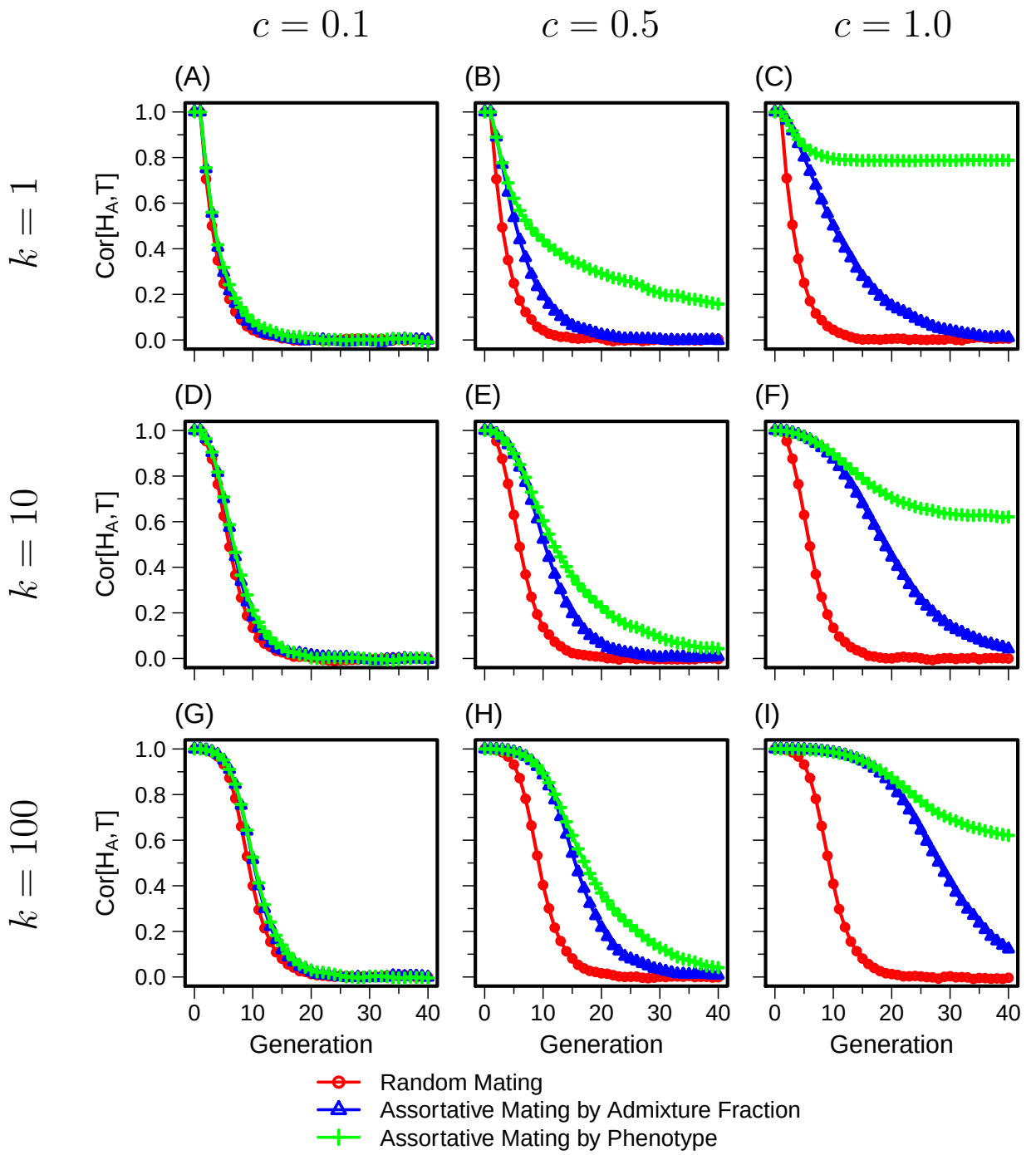

Figure 4: Correlation between admixture fraction and quantitative trait value ($\mathrm{Cor}[H_A, T]$) as a function of time. All parameter values in panel (E) follow the base case described in Section 3.2; the number of quantitative trait loci k and the assortative mating strength c vary across panels. In each panel, for a given (k, c) pair, for each mating scheme, the mean of 100 simulated trajectories is plotted. The red, blue, and green curves represent results from random mating, assortative mating by admixture fraction, and assortative mating by phenotype, respectively. (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) $k = 100$, $c = 1.0$.

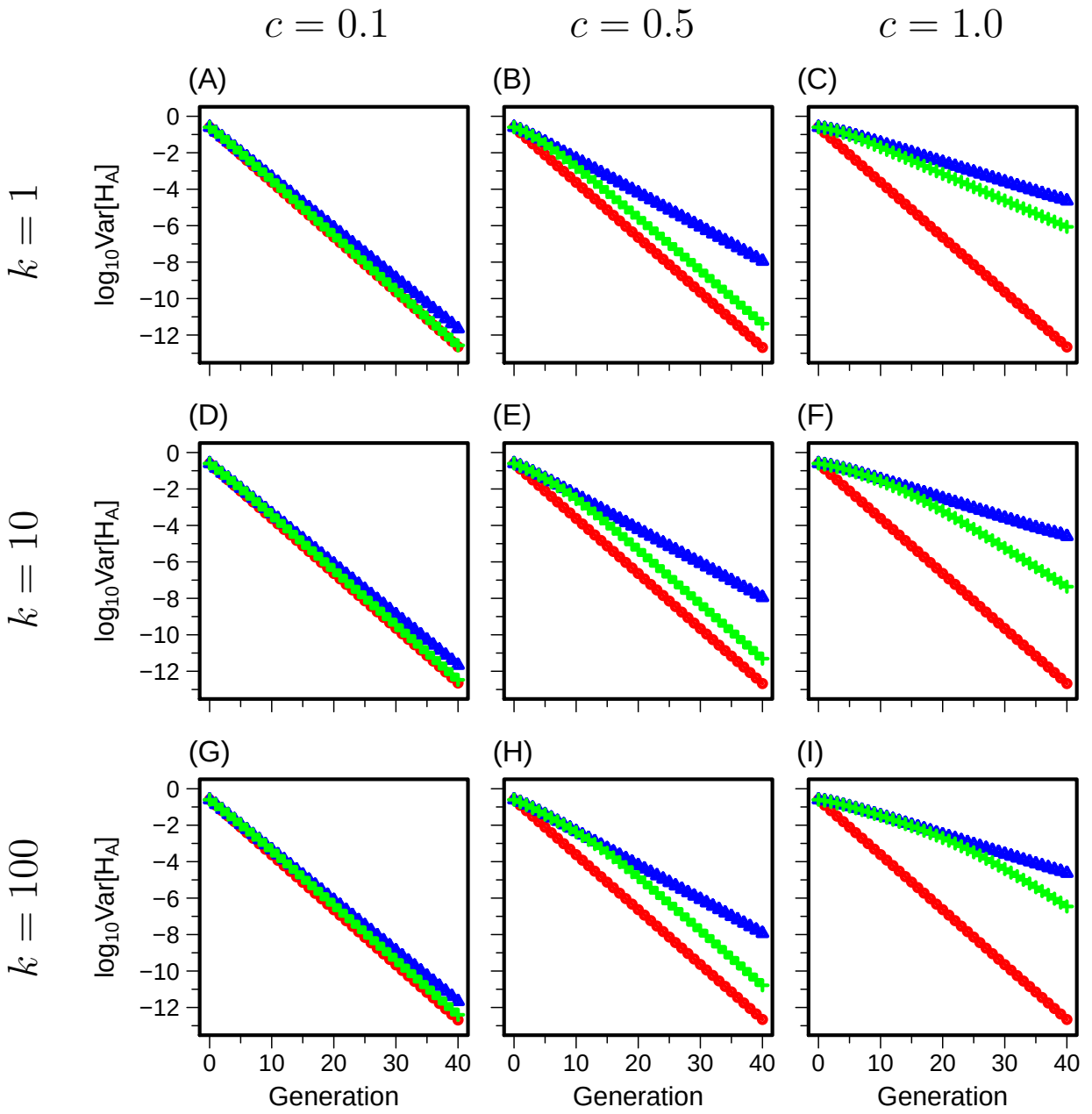

Figure 5: Variance of admixture fraction ($\mathrm{Var}[H_A]$) as a function of time. The simulations shown are the same ones from Figure 4. (A) $k = 1$, $c = 0.1$. (B) $k = 1$, $c = 0.5$. (C) $k = 1$, $c = 1.0$. (D) $k = 10$, $c = 0.1$. (E) $k = 10$, $c = 0.5$. (F) $k = 10$, $c = 1.0$. (G) $k = 100$, $c = 0.1$. (H) $k = 100$, $c = 0.5$. (I) $k = 100$, $c = 1.0$. Colors and symbols follow Figure 4. The y-axis is plotted on a logarithmic scale.

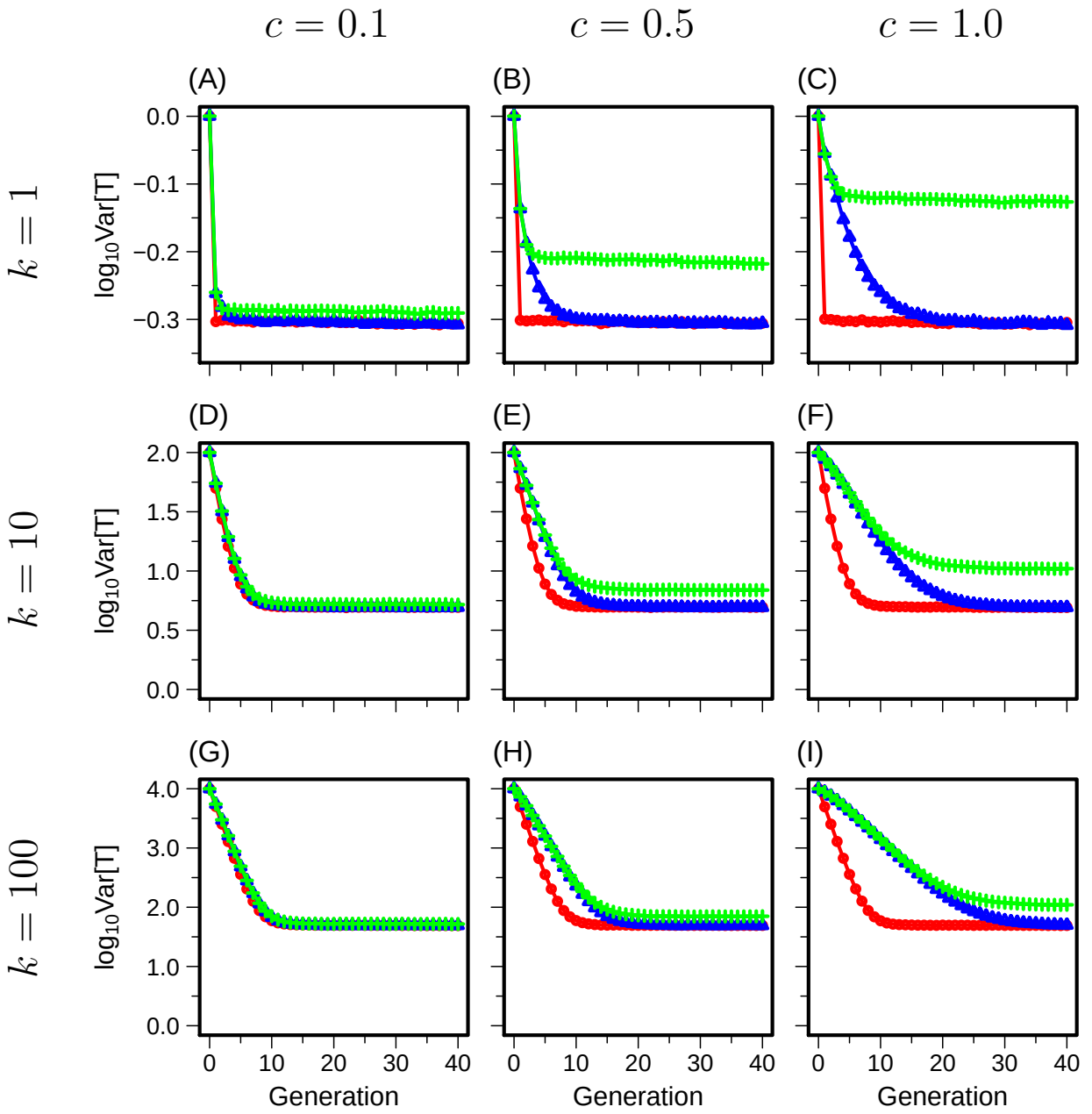

Figure 6: Variance of the phenotype (Var[T]) as a function of time. The simulations shown are the same ones from Fig. 4. (A) $k = 1$, $c = 0.1$. (B) $k = 1$, $c = 0.5$. (C) $k = 1$, $c = 1.0$. (D) $k = 10$, $c = 0.1$. (E) $k = 10$, $c = 0.5$. (F) $k = 10$, $c = 1.0$. (G) $k = 100$, $c = 0.1$. (H) $k = 100$, $c = 0.5$. (I) $k = 100$, $c = 1.0$. Colors and symbols follow Figure 4. The y-axis is plotted on a logarithmic scale.

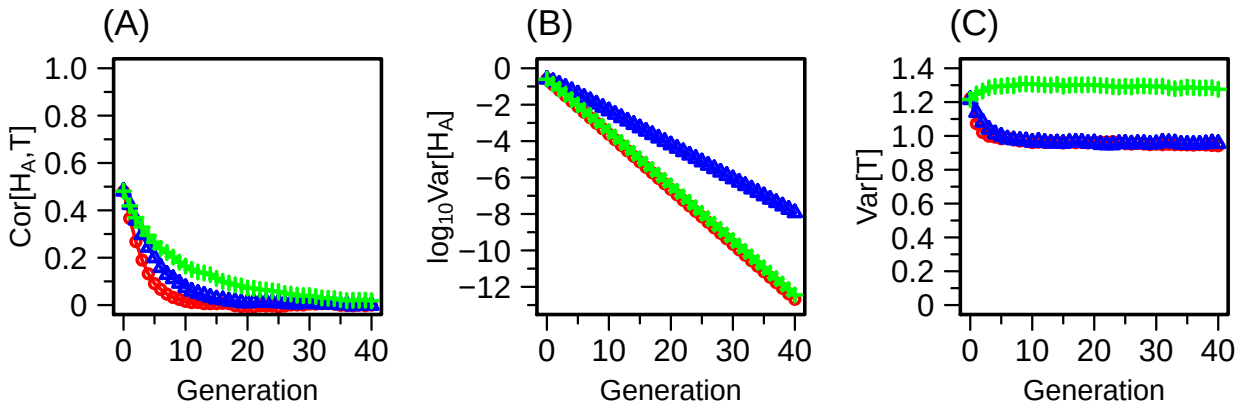

Figure 7: $\mathrm{Cor}[H_A, T]$, $\mathrm{Var}[H_A]$, and $\mathrm{Var}[T]$ in a model in which the allele frequencies p_i and q_i in the source populations S_1 and S_2 are drawn from a simulation rather than being treated as fixed at 1 and 0, respectively. All other parameters are kept at the values of the base case (Section 3.2). (A) Correlation between admixture fraction and trait ($\mathrm{Cor}[H_A, T]$). (B) Variance of the admixture fraction ($\mathrm{Var}[H_A]$). (C) Variance of the phenotype ($\mathrm{Var}[T]$). 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 Section 3.1. 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 [22]. The y-axis of $\mathrm{Var}[H_A]$ is plotted on a logarithmic scale. Results using other replicates of simulated allele frequencies with $k = 10$ are shown in Figure S4.

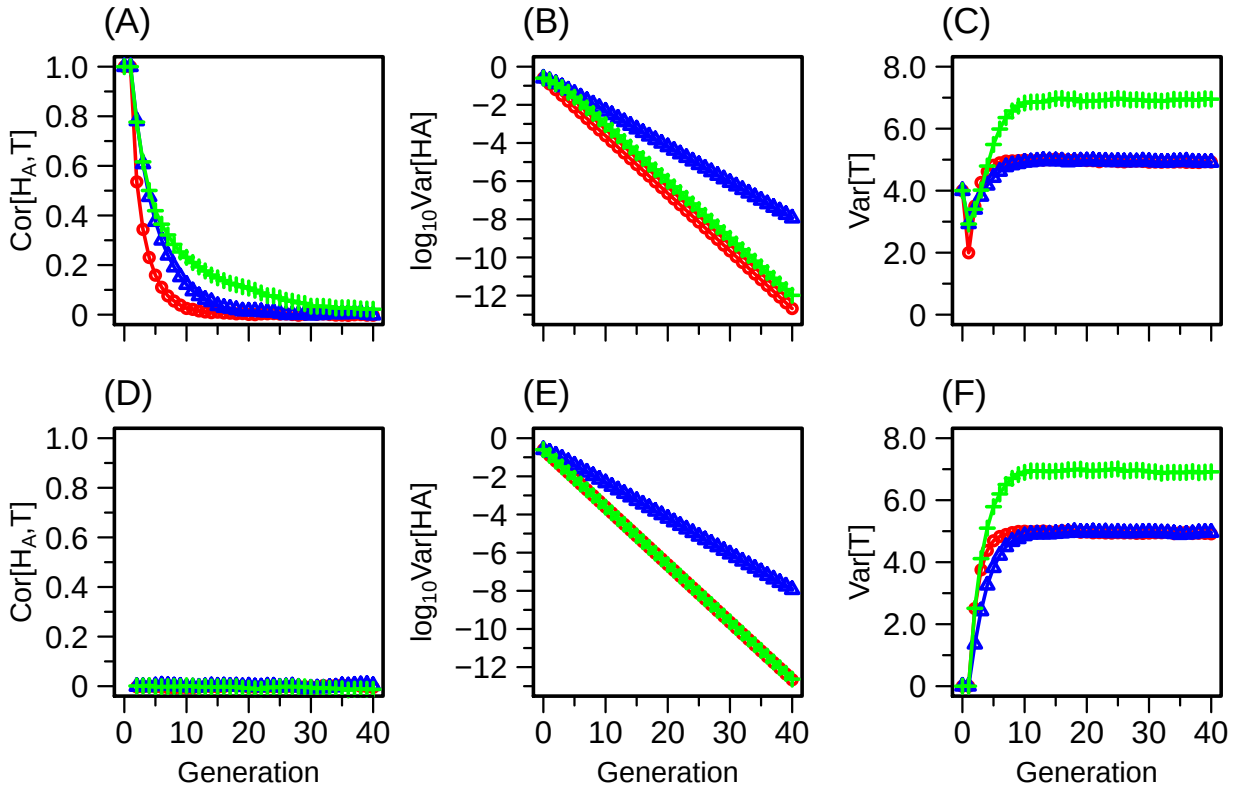

Figure 8: $\mathrm{Cor}[H_A, T]$, $\mathrm{Var}[H_A]$, and $\mathrm{Var}[T]$ under a trait model in which trait loci do not systematically have greater values in one source population: $P(X_i = 1) = P(X_i = 0) = 0.5$. All other parameters are kept at the values of the base case (Section 3.2). Of $k = 10$ trait loci, we denote the number of randomly selected loci to have $X_i = 1$ by z. (A) $\mathrm{Cor}[H_A, T]$, $z = 6$. (B) $\mathrm{Var}[H_A]$, $z = 6$. (C) $\mathrm{Var}[T]$, $z = 6$. (D) $\mathrm{Cor}[H_A, T]$, $z = 5$. (E) $\mathrm{Var}[H_A]$, $z = 5$. (F) $\mathrm{Var}[T]$, $z = 5$. 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 $\mathrm{Var}[H_A]$ is plotted on a logarithmic scale. Results using other replicates of simulated allele frequencies with $k = 10$ are shown in Figures S5 (for $z = 6$) and S6 (for $z = 5$).

Supplementary Material

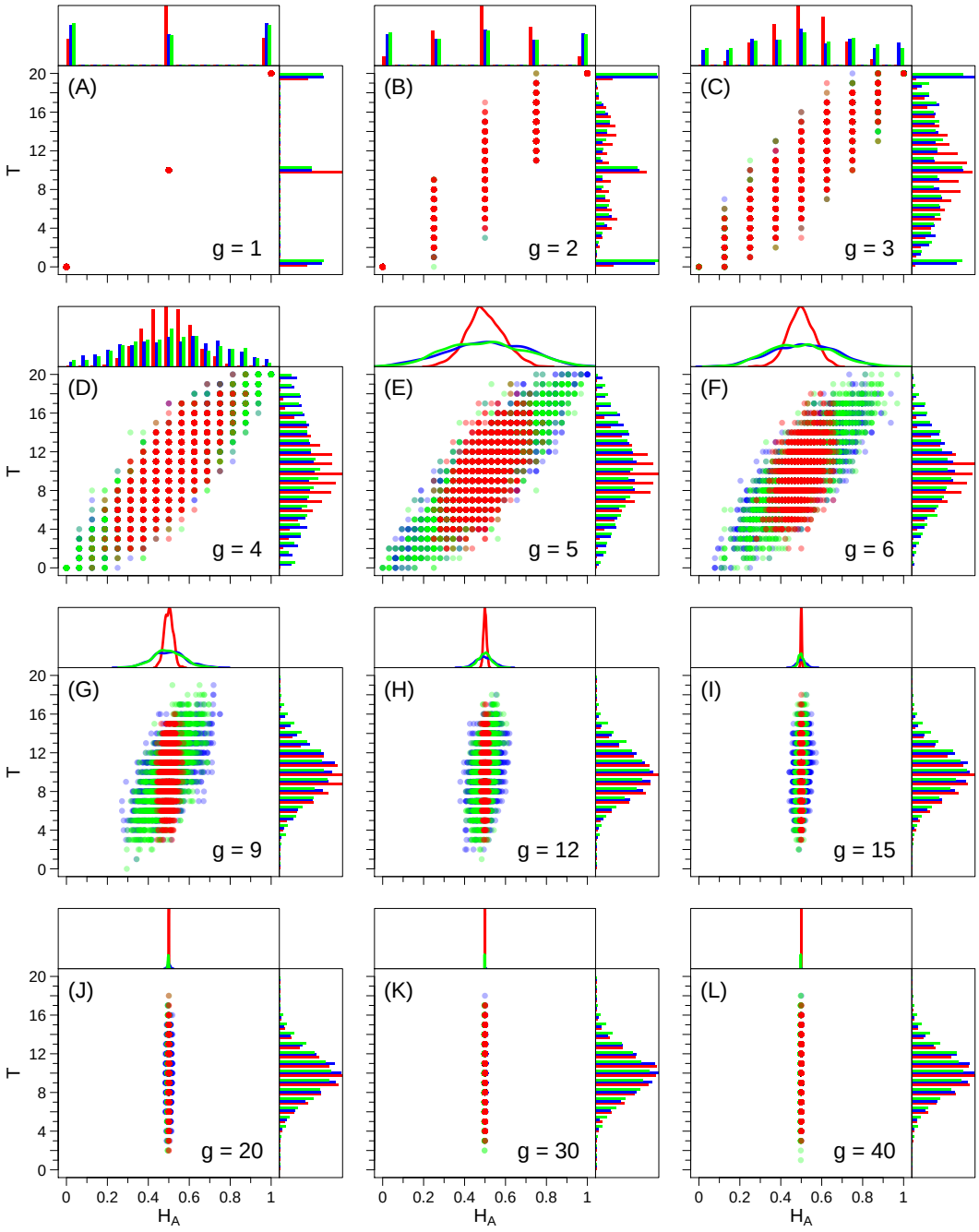

Figure S1: Joint distribution of H_A and T as a function of time. The simulation shown is the same one from Figures 4E, 5E, and 6E, using the parameters from base case (Section 3.2). As described in Section 2.1 and 2.2, the possible values for the admixture fraction at generation g are $0, 1/2^g, 2/2^g, \ldots, (2^g-1)/2^g, 1$, 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.

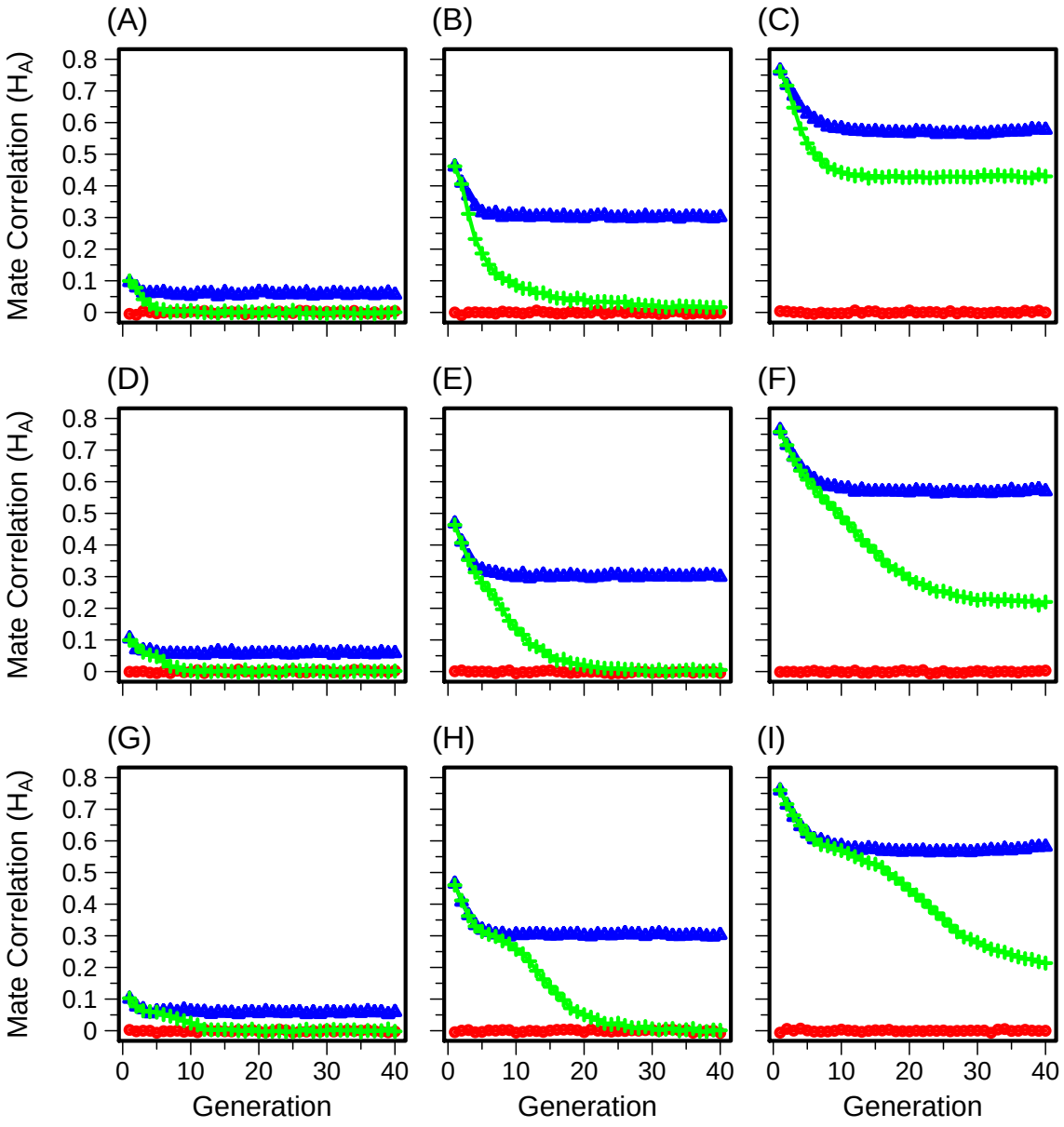

Figure S2: The correlation coefficient $\mathrm{Cor}[H^f_{A,g}, H^m_{A,g}]$ between the admixture fractions of the members of mating pairs as a function of time. The simulations shown are the same ones from Fig. 4. (A) $k = 1$, $c = 0.1$. (B) $k = 1$, $c = 0.5$. (C) $k = 1$, $c = 1.0$. (D) $k = 10$,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

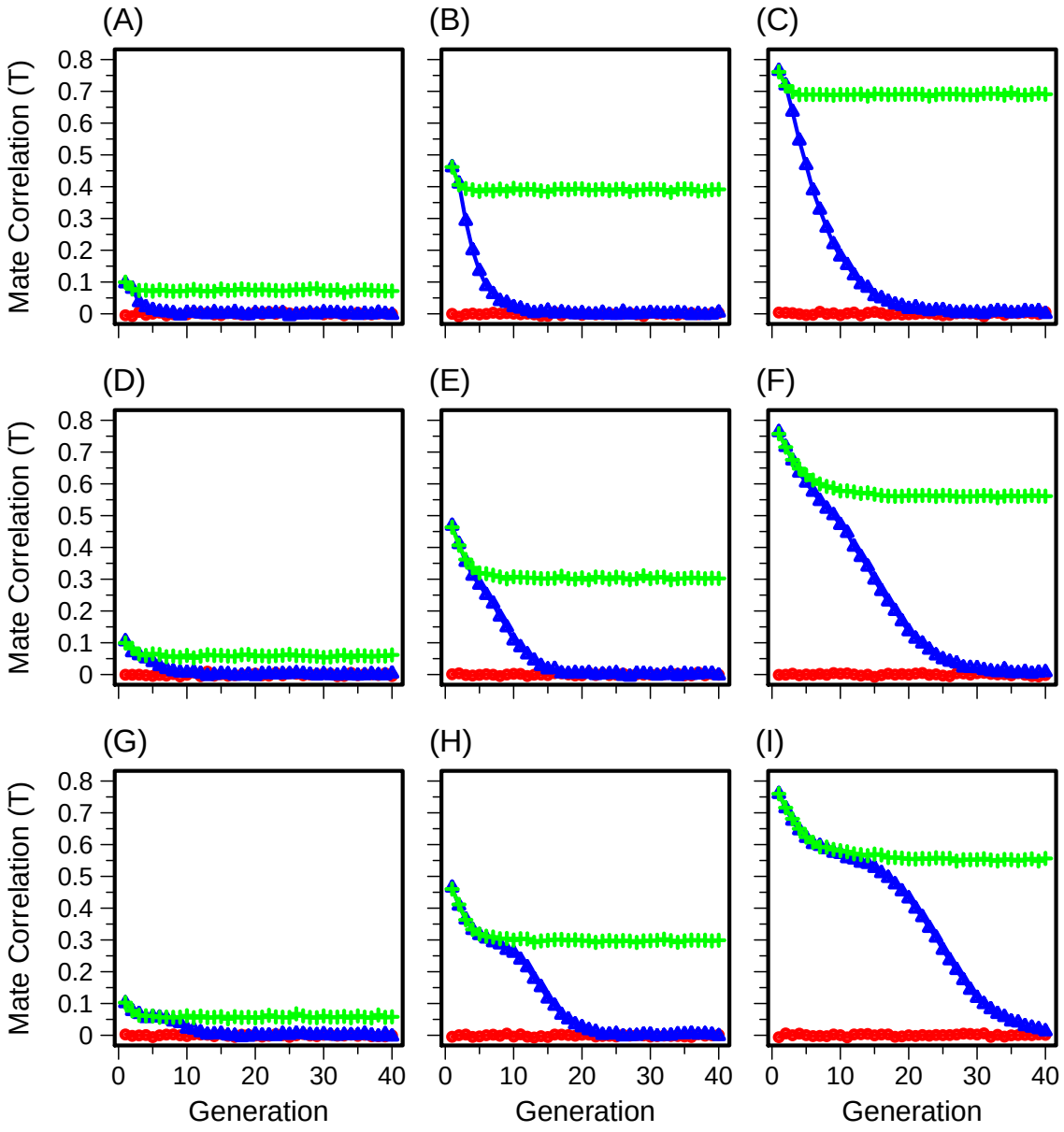

Figure S3: The correlation coefficient $\mathrm{Cor}[T_g^f, T_g^m]$ between the phenotypes of the members of mating pairs as a function of time. The simulations shown are the same ones from Fig. 4. (A) $k = 1$, $c = 0.1$. (B) $k = 1$, $c = 0.5$. (C) $k = 1$, $c = 1.0$. (D) $k = 10$, $c = 0.1$. (E) $k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

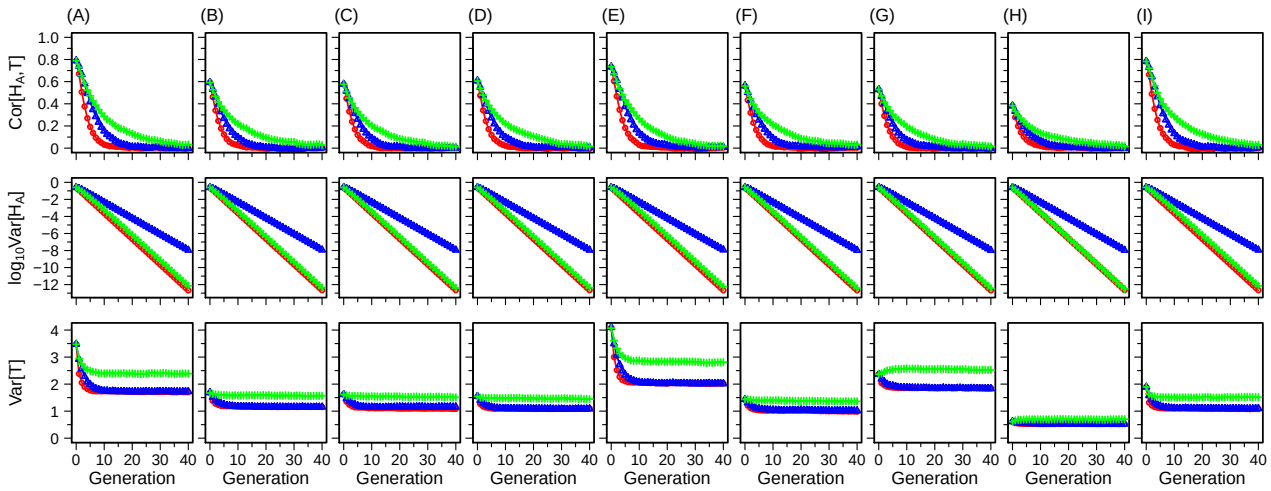

Figure S4: Cor$[H_A, T]$ (top row), Var$[H_A]$ (middle row), and Var$[T]$ (bottom row) using different replicate sets of simulated allele frequencies p_i and q_i with $k = 10$ loci, as described in Section 3.1 and Figure 7. Different columns represent results from different replicates of simulated p_i 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.

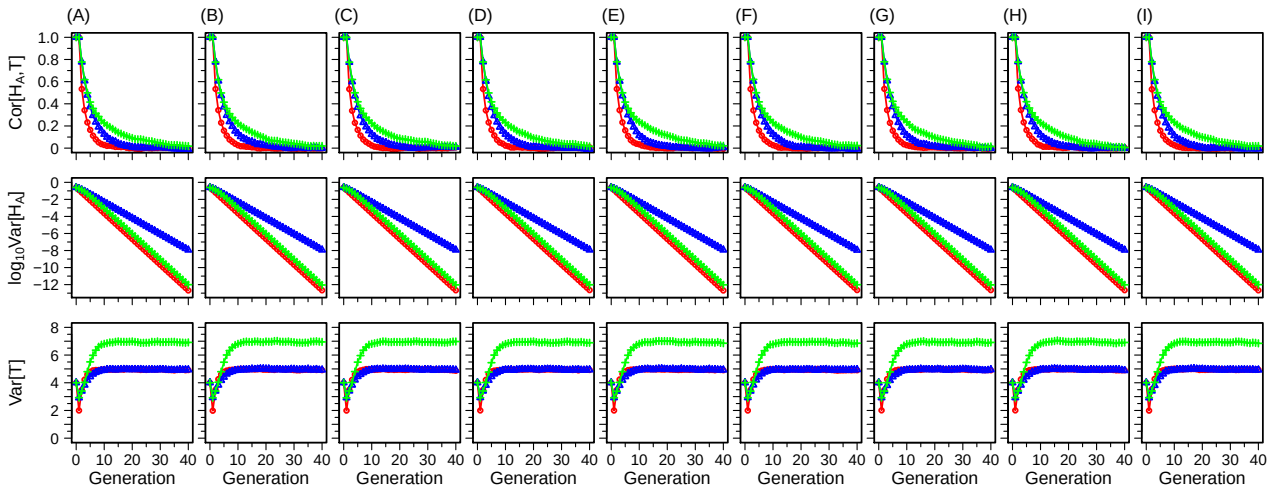

Figure S5: $\mathrm{Cor}[H_A, T]$ (top row), $\mathrm{Var}[H_A]$ (middle row), and $\mathrm{Var}[T]$ (bottom row) using different replicates of a set of X_i for $k = 10$ loci sampled from a Binomial(10, $\frac{1}{2}$) distribution with a constraint $z = 6$ (Section 2.2 and Figure 8A-C). Colors and symbols follow Figure 4. The y-axis of $\mathrm{Var}[H_A]$ is plotted on a logarithmic scale.

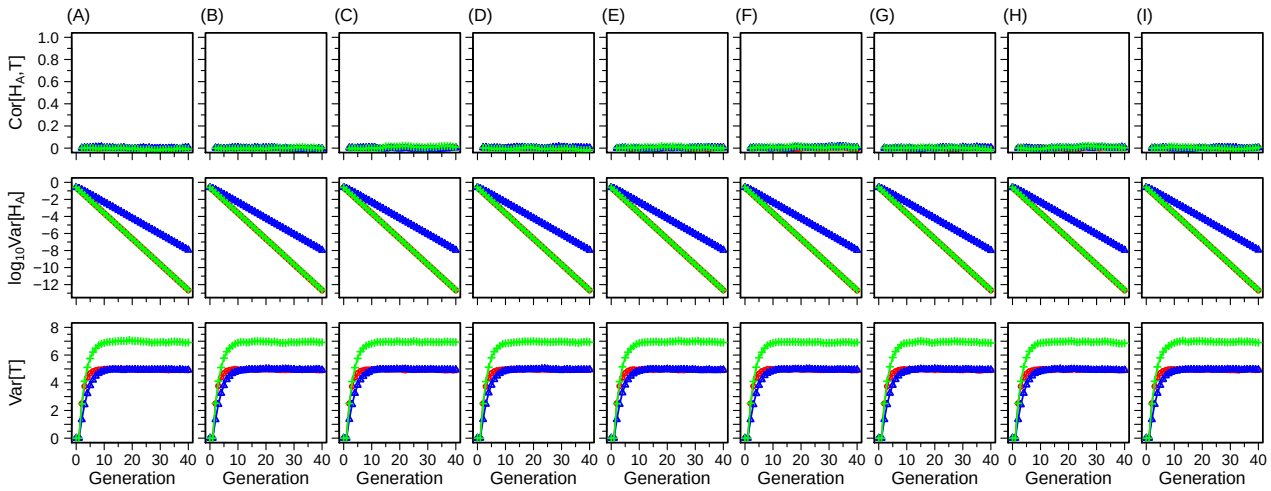

Figure S6: Cor$[H_A, T]$ (top row), Var$[H_A]$ (middle row), and Var$[T]$ (bottom row) using different replicates of a set of X_i for $k = 10$ loci sampled from a Binomial(10, $\frac{1}{2}$) distribution with a constraint $z = 5$ (Section 2.2 and Figure 8D-F). Colors and symbols follow Figure 4. The y-axis of Var$[H_A]$ is plotted on a logarithmic scale.