

# Genetic nurturing, missing heritability, and causal analysis in genetic statistics

Hao Shen<sup>a</sup> and Marcus W. Feldman<sup>a,1</sup>

<sup>a</sup>Department of Biology, Stanford University, Stanford, CA 94305

Contributed by Marcus W. Feldman, August 25, 2020 (sent for review July 28, 2020; reviewed by Aviv Bergman and Bruce S. Weir)

**Genetic nurturing, the effect of parents' genotypes on offspring phenotypes through parental phenotypic transmission, can be modeled in terms of gene–culture interactions. This paper first uses a simple one-locus, two-phenotype gene–culture cotransmission model to compute the effect of genetic nurturing in terms of regression of children's phenotypes on transmitted and nontransmitted alleles. With genetic nurturing, interpreting heritability and hence the meaning of “missing heritability” becomes problematic. Other factors, for example, population subdivision and assortative mating, generate similar signals to those of genetic nurturing, namely, correlation between parents' nontransmitted alleles and children's phenotypes. Corrections must be made for these to isolate the signal of genetic nurturing. Finally, a unified causal framework is constructed for genetic nurturing, population subdivision, and assortative mating. Causal and noncausal paths from transmitted and nontransmitted alleles to children's phenotypes are identified and investigated in the presence of genetic nurturing, population subdivision, and assortative mating. Using causal analysis, assumptions made in inferring direct and indirect effects are then clarified and evaluated in a broader causal context.**

genetic nurturing | cultural transmission | missing heritability | population subdivision and assortative mating | correlation and causality

The nature vs. nurture problem is fundamental to most human sciences. History is replete with disastrous human suffering caused by malevolent use of positions taken regarding this problem. In the second half of the 20th century, after the modern evolutionary synthesis, researchers in human behavior started to incorporate simple models from statistical genetics, in the process extending such concepts as heritability beyond their original function. The notion of heritability was first introduced by Lush (1) in 1937 in animal breeding to predict the effectiveness of artificial selection. From the 1960s, people such as Arthur Jensen (2) and others used an estimate known as “broad sense heritability” as a statistical measure of genetic determination in research into human behavior. Their high heritability estimates for IQ led to a genetic deterministic view of human intelligence (3, 4). These estimates were often based on studies of twins reared together and apart (5) or on other correlations between relatives. However, such estimates of broad heritability cannot disentangle gene-by-environment interactions from purely genetic effects (6–9).

The debates of the 1970s led Cavalli-Sforza and Feldman (10) to construct generative models of cultural and genotypic transmission from which one could assess the contribution of vertical cultural transmission to phenotypic variation. They showed that such cultural transmission could lead to inflated estimates of genetic heritability.

Correlations between relatives continued to be the primary tool for estimation of the genetic contribution to phenotypic variance throughout the 1970s and 1980s (3, 4, 11, 12), although the importance of cultural transmission in inflating such correlations was recognized by several investigators (7, 13–17). Analysis of the model constructed by Cavalli-Sforza and Feldman (10) showed that parents' phenotypes, if transmitted through vertical cultural

transmission, could have strong effects on correlations between relatives. This approach was used, and the results confirmed, by Cloninger et al. (13–15) and also by Morton and coworkers in Hawaii (17).

In the past 10 y, with the growing popularity of genome-wide association studies (GWAS), heritability has been estimated from data by accumulating the variance explained by millions of single-nucleotide polymorphisms (SNPs) and in general SNP heritability is lower than earlier estimates based on correlations between relatives. This inconsistency is known as “missing heritability” (18, 19) and represents an important conundrum in human statistical genetics. Different factors can contribute to this inconsistency and we divide them into two categories: genetic factors and demographic factors. Genetic factors include too few SNPs (20–22) and complex genetic interactions [linkage disequilibrium, epistasis (23), etc.]. Demographic factors include cultural transmission (19, 22, 24), population subdivision (25), and assortative mating (24). Both categories can contribute to missing heritability and their relative importance can vary from trait to trait. Historically, barring a few exceptions, researchers have focused primarily on genetic factors.

Recently, Kong et al. (26) found substantial correlation between parents' nontransmitted alleles and children's phenotypes. They suggested that this correlation could be generated by the effect of parents' phenotypes, which are influenced by parents' genotypes, on children's phenotypes. This sequence, from parents' genotypes to parents' phenotypes to children's

## Significance

**Correlation between genotypes and phenotypes can be produced by genetic nurturing, namely the effect of parents' genotypes on their offspring's phenotypes through the parents' phenotypes. Population subdivision and assortative mating can give rise to correlations between genotypes and phenotypes similar to those due to genetic nurturing. Variances and correlations may not reveal causal relationships in the presence of these complexities. We analyze mechanistic models of genetic nurturing, population subdivision, and assortative mating and compare these with results obtained within the framework of modern causal analysis. Our results clarify statistical signals emanating from correlations between nontransmitted alleles and offspring phenotypes and reveal difficulties with standard linear models in the interpretation of heritability, in particular, the concept of missing heritability.**

Author contributions: H.S. and M.W.F. designed research, performed research, and wrote the paper.

Reviewers: A.B., Albert Einstein College of Medicine; and B.S.W., University of Washington.

The authors declare no competing interest.

This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

<sup>1</sup>To whom correspondence may be addressed. Email: [mfeldman@stanford.edu](mailto:mfeldman@stanford.edu).

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2015869117/-DCSupplemental>.

First published September 28, 2020.

phenotypes, was called “genetic nurturing.” Subsequent studies connected genetic nurturing to the missing heritability problem (27–29). These studies appear to be closely related to the original model of Cavalli-Sforza and Feldman (10) and its generalization by Feldman et al. (8). Here we propose a simple extension of these models and show how vertical cultural transmission can lead to genetic nurturing. This affects the linear regressions that have been used to separate the contributions to phenotypes by nature and nurture, and we show how it affects our understanding of the missing heritability problem. We also explore how population subdivision and assortative mating, which are common for many traits in humans (30, 31), can generate correlation between nontransmitted alleles and children’s phenotype. A natural next step is to analyze how these different factors could produce such a correlation in a causal framework, which can help expose the nature–nurture conundrum (32–35), in the presence of complex population dynamics and possible confounding factors. Thus, we propose a unified causal framework for genetic nurturing, population subdivision, and assortative mating, which extends to nonlinear cases such as the path analysis approach used by Kong et al. (26). We also examine the approach they used to estimate direct genetic effects and make corrections for assortative mating.

### A Model with Cultural Transmission

Genetic nurturing (26) entails that parents’ genotypes can have an indirect effect on children’s phenotypes by influencing parents’ phenotypes. This implies that children’s phenotypes will be determined by both children’s genotypes and parents’ phenotypes, where the influence of parents’ phenotypes occurs via vertical cultural transmission. Cavalli-Sforza and Feldman (10) proposed a quantitative genetic model incorporating vertical cultural transmission and showed that the heritability could be overestimated if there was vertical cultural transmission. However, for quantitative traits, it is very difficult to analyze the mathematical dynamics of a trait’s distribution. To address this issue and develop some theoretical understanding of the genetic nurturing effect, we construct a discrete trait model based on Cavalli-Sforza and Feldman (10) and show how the genetic nurturing effect can be included in the model.

We consider one diploid locus with two alleles,  $A_1$  and  $A_2$ . Each genotype can be one of two phenotypic variants represented by bar and nonbar. Thus the six pheno-genotypes can be written as  $\overline{A_i A_j}$  and  $A_i A_j$  with  $i, j = 1, 2$ . The probability of a child being bar is determined by the parents’ phenotypes and the child’s genotype. To be specific, denote by  $X_{ch}$  the phenotypic state of the child, which is 1 for bar and 0 for nonbar;  $G_{ch}$  the genotypic state of the child, which is 1 for  $A_1 A_1$ , 2 for  $A_1 A_2$ , and 3 for  $A_2 A_2$ ;  $\varepsilon_{G_{ch}}$  the contribution of the child’s genotype when its genotype is  $G_{ch}$ ; and  $g(y, x)$  the contribution of a parent’s phenotype when the parent’s phenotype is  $x$  and the child’s genotype is  $y$ . Then we have

$$P(X_{ch} = 1 | G_{ch}, X_p, X_m) = \varepsilon_{G_{ch}} + g(G_{ch}, X_p) + g(G_{ch}, X_m), \quad [1]$$

where  $g(y, x) = a_y I(x = 1) + b_y I(x = 0)$ , with  $I(\cdot)$  the index function, and  $X_p$  and  $X_m$  stand for the phenotypic states of the father and mother, respectively.

Since we can always move the common part of  $a$  and  $b$  into  $\varepsilon$ , we reduce the number of parameters by specifying for  $y = 1, 2, 3$ ,

$$a_y - b_y = e_y; 2b_y + \varepsilon_y = c_y. \quad [2]$$

Table 1 describes this model, where

$$P(X_{ch} = 1 | G_{ch}, X_p, X_m) = c_{G_{ch}} + e_{G_{ch}} [I(X_p = 1) + I(X_m = 1)]. \quad [3]$$

We denote the frequency of the bar phenotype with genotype  $y$  at generation  $n$  by  $x_y^n$ , the allele frequency of  $A_1$  by  $p$ , and the allele frequency of  $A_2$  by  $q = 1 - p$ . Without selection and assortative mating, the population is in Hardy–Weinberg equilibrium and  $p$  does not change. The frequencies of the six pheno-genotypes  $\overline{A_1 A_1}, A_1 A_1, \overline{A_1 A_2}, A_1 A_2, \overline{A_2 A_2}, A_2 A_2$  are  $p^2 x_1, p^2(1 - x_1), 2pqx_2, 2pq(1 - x_2), q^2 x_3, q^2(1 - x_3)$ , respectively.

Now consider the evolution under random mating. On the one hand, the probability of being bar for  $A_1 A_1$  in generation  $n + 1$  is  $x_1^{n+1}$ . On the other hand, it equals  $c_1 + 2e_1 P_1^n$ , where  $P_1^n$  represents the probability that one of the parents (say, the mother) in generation  $n$  has the bar phenotype given that the child’s pheno-genotype is  $\overline{A_1 A_1}$ . Since the child’s genotype is  $A_1 A_1$ , the mother’s genotype has to be  $A_1 A_i$ , where allele  $i$  has probability  $p$  being 1 and probability  $q$  being 2. Since the probabilities of being bar for  $A_1 A_1$  and  $A_1 A_2$  in generation  $n$  are  $x_1^n$  and  $x_2^n$ , we have  $P_1^n = px_1^n + qx_2^n$ . Thus,  $x_1^{n+1} = 2e_1 px_1^n + 2e_1 qx_2^n + c_1$ . Corresponding recursions for  $x_2$  and  $x_3$  can be derived, and we have

$$\begin{aligned} x_1^{n+1} &= 2e_1 px_1^n + 2e_1 qx_2^n + c_1; \\ x_2^{n+1} &= e_2 px_1^n + e_2 x_2^n + e_2 qx_3^n + c_2; \\ x_3^{n+1} &= 2e_3 qx_3^n + 2e_3 px_2^n + c_3. \end{aligned} \quad [4]$$

The globally stable equilibrium of Eq. 4 (SI Appendix, section A), denoted by  $(x_1, x_2, x_3)$ , is  $(x_1, x_2, x_3) = (\frac{B_1}{A}, \frac{B_2}{A}, \frac{B_3}{A})$  with

$$\begin{aligned} A &= 1 - e_2 - 2e_1 p - 2e_3 q + 2e_1 e_2 p^2 + 2e_2 e_3 q^2 \\ &\quad + 4e_1 e_3 pq; \\ B_1 &= (1 - e_2 - 2e_3 q + 2e_2 e_3 q^2) c_1 + 2e_1 q(1 - 2e_3 q) c_2 \\ &\quad + 2e_1 e_2 q^2 c_3; \\ B_2 &= e_2 p(1 - 2e_3 q) c_1 + (1 - 2e_3 q)(1 - 2e_1 p) c_2 \\ &\quad + e_2 q(1 - 2e_1 p) c_3; \\ B_3 &= 2e_3 e_2 p^2 c_1 + 2e_3 p(1 - 2e_1 p) c_2 + (1 - e_2 - 2e_1 p \\ &\quad + 2e_1 e_2 p^2) c_3. \end{aligned} \quad [5]$$

### Regression Analysis

We regress the child’s phenotype  $X_{ch}$  on alleles transmitted from the parents to the child and alleles not transmitted to the child.

**Table 1. Transmission table**

	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$
$\overline{A_i A_j} \times \overline{A_k A_h}$	$\varepsilon_1 + 2a_1 = c_1 + 2e_1$	$\varepsilon_2 + 2a_2 = c_2 + 2e_2$	$\varepsilon_3 + 2a_3 = c_3 + 2e_3$
$\overline{A_i A_j} \times A_k A_h$	$\varepsilon_1 + a_1 + b_1 = c_1 + e_1$	$\varepsilon_2 + a_2 + b_2 = c_2 + e_2$	$\varepsilon_3 + a_3 + b_3 = c_3 + e_3$
$A_i A_j \times \overline{A_k A_h}$	$\varepsilon_1 + a_1 + b_1 = c_1 + e_1$	$\varepsilon_2 + a_2 + b_2 = c_2 + e_2$	$\varepsilon_3 + a_3 + b_3 = c_3 + e_3$
$A_i A_j \times A_k A_h$	$\varepsilon_1 + 2b_1 = c_1$	$\varepsilon_2 + 2b_2 = c_2$	$\varepsilon_3 + 2b_3 = c_3$

The first row shows the child’s genotype; the first column represents parents’ pheno-genotypes, and the entries represent the child’s probability of having the bar phenotype conditioned on child’s genotype (first row) and parents’ phenotypes (first column).

Let the number of  $A_2$  in parental alleles transmitted to the child be  $N_{pa}^t \in \{0, 1, 2\}$  and the number of  $A_2$  in parental alleles not transmitted to the child be  $N_{pa}^n \in \{0, 1, 2\}$ . Then, since  $N_{pa}^t$  and  $N_{pa}^n$  are independent random variables, we can write the regression equation

$$X_{ch} = \beta_1 N_{pa}^t + \beta_2 N_{pa}^n + \delta, \quad [6]$$

where  $\beta_1 = \frac{Cov(X_{ch}, N_{pa}^t)}{Var(N_{pa}^t)}$ ;  $\beta_2 = \frac{Cov(X_{ch}, N_{pa}^n)}{Var(N_{pa}^n)}$ , and  $\delta$  is the sum of the intercept and the residual. Calculation of these expressions is shown in [SI Appendix, section B](#), and we obtain

$$Var(N_{pa}^t) = 2pq; \quad [7a]$$

$$Var(N_{pa}^n) = 2pq; \quad [7b]$$

$$Cov(X_{ch}, N_{pa}^t) = 2pq[p(x_2 - x_1) + q(x_3 - x_2)]; \quad [7c]$$

$$Cov(X_{ch}, N_{pa}^n) = 2pq[p(pe_1 + qe_2)(x_2 - x_1) + q(pe_2 + qe_3)(x_3 - x_2)]. \quad [7d]$$

Then the coefficients in Eq. 6 are

$$\beta_1 = \frac{Cov(X_{ch}, N_{pa}^t)}{Var(N_{pa}^t)} = p(x_2 - x_1) + q(x_3 - x_2); \quad [8a]$$

$$\beta_2 = \frac{Cov(X_{ch}, N_{pa}^n)}{Var(N_{pa}^n)} = p(pe_1 + qe_2)(x_2 - x_1) + q(pe_2 + qe_3)(x_3 - x_2). \quad [8b]$$

If the transmission process is purely genetic, that is,  $e_s$  are zero, we have  $\beta_2 = 0$ , and  $\beta_1 = p(c_2 - c_1) = q(c_3 - c_2)$ , which corresponds to the classical average effect in Falconer and Mackay (36) and the regression slope in Lynch and Walsh (ref. 37, pp. 65–67). With cultural transmission, i.e., nonzero  $e_s$ ,  $\beta_1$  still has the familiar form [8a](#) due to independence of transmitted and nontransmitted alleles. However, the frequencies  $x_1, x_2, x_3$  will be complicated expressions in  $e_s$  and  $c_s$  shown in Eq. 5, and  $\beta_2$  will not be zero. We can understand this result more clearly from the way parents' nontransmitted alleles influence children's phenotypes. The nontransmitted alleles are included in parents' genotypes, which affect parents' phenotypes, and thus influence children's phenotypes by cultural transmission ( $e_s$ ). When the cultural transmission pathway (genetic nurturing pathway) is absent ( $e_s = 0$ ), there is no influence of nontransmitted alleles.

Kong et al. (26) assumed that both the transmitted and nontransmitted alleles participate in genetic nurturing, which is also the case in our model. Since the transmitted alleles influence the parents' phenotypes, as do the nontransmitted alleles, they calculated the direct effect [ $\delta$  in Kong et al.'s (26) paper] by subtracting the regression coefficients of nontransmitted alleles from those of the transmitted alleles. In our model, the corresponding quantity is  $\beta_1 - \beta_2$ , which represents the direct genetic effect when the parents' phenotypes have the same effect on different children's genotypes. To show this, we assume  $c_s$  are linear, i.e.,  $(c_1, c_2, c_3) = (c, c + d, c + 2d)$ , and  $e_s$  are the same, i.e.,  $(e_1, e_2, e_3) = (e, e, e)$ . Using Eqs. 8 and 5 we then have

$$\beta_1 = p(x_2 - x_1) + q(x_3 - x_2) = \frac{d}{1 - e} \quad [9a]$$

and

$$\beta_2 = e\beta_1 = \frac{ed}{1 - e}, \quad [9b]$$

which implies

$$\beta_1 - \beta_2 = (1 - e)\beta_1 = d. \quad [9c]$$

Although  $\beta_1$  and  $\beta_2$  do not constitute “total” and “indirect” effects due to the confounding of grandparents' phenotypic effects, their difference  $d$  is exactly the direct genetic effect. This analysis can be extended to cases with arbitrary  $c_s$  and the same  $e$ , but it fails when  $e_s$  are different. This is because the genetic nurturing influences the regression in two ways: First, as shown above, it incorporates the nontransmitted alleles. Second, it influences the equilibrium frequency of the bar phenotype for each genotype, which makes it difficult to extract the direct genetic effect, as claimed by Kong et al. (26). (For more detailed analysis, see [A Unified Causal Framework for Genetic Nurturing, Population Subdivision, and Assortative Mating](#) and [SI Appendix, sections E and J](#).) We will see in the next part that this observation also leads to an interesting view of the missing heritability problem.

### Missing Heritability?

The missing heritability problem (18) arises when the heritability calculated from regression on alleles is lower than the heritability calculated from correlations between relatives. As pointed out by Cavalli-Sforza and Feldman (10), the latter heritability can be overestimated if there is vertical cultural transmission. To quantify this in our model, we regress the child's phenotype on the parental phenotypes. The parent–offspring phenotype covariance ([SI Appendix, section C](#)) can be written

$$Cov(X_{ch}, X_p + X_m) = \Delta_1 + \Delta_2, \quad [10]$$

where

$$\Delta_1 = 2pq\beta_1^2; \quad [11a]$$

$$\Delta_2 = 2p(pe_1 + qe_2)(px_1 + qx_2)[1 - (px_1 + qx_2)] + 2q(pe_2 + qe_3)(px_2 + qx_3)[1 - (px_2 + qx_3)]. \quad [11b]$$

(Note that for convenience, we use the sum of the parental phenotypic values rather than their average.) Note that the additive genetic variance calculated by regression on transmitted alleles is  $V_A = Var(N_{pa}^t)\beta_1^2 = 2pq\beta_1^2 = \Delta_1$ , and  $\Delta_2$  is always positive. Thus the missing heritability can be explained qualitatively by  $Cov(X_{ch}, X_p + X_m) - V_A = \Delta_2 > 0$ . More formally, let  $j_1 = px_1 + qx_2$ ;  $j_2 = px_2 + qx_3$ , which gives

$$\frac{\Delta_1}{\Delta_2} = \frac{pq(j_1 - j_2)^2}{p(pe_1 + qe_2)j_1(1 - j_1) + q(pe_2 + qe_3)j_2(1 - j_2)} \quad [12]$$

In the parameter region where  $\frac{\Delta_1}{\Delta_2} \ll 1$ , the heritability estimated from regression on alleles is much smaller than the heritability estimated from correlations between relatives. This seems to explain the missing heritability problem. However, our analysis raises two problems in interpreting missing heritability, which render it meaningless in the context of cultural transmission. First, the genetic nurturing effect entails that the additive genetic variance computed from GWAS and that computed from correlations between relatives both depend on the equilibrium frequencies ( $x_i$ s), which depend on both  $c_s$  and  $e_s$ . This is true even in the case where  $e_s$  are all the same and  $c_s$  are linear. In this case, from Eq. 9a, we have  $\Delta_1 = \frac{2pqd^2}{(1 - e)^2}$ , and calculation of  $Cov(X_{ch}, X_p + X_m)$  shows that this also depends on  $d$  and  $e$ . (For details, see [SI Appendix, section D](#).) Thus the so-called “missing heritability problem” is generated by comparing two quantities, neither of which is purely genetic. Second, although cultural transmission can contribute to missing heritability, the absence of missing heritability does not imply that cultural transmission is weak or absent. There exist parameter regions where cultural transmission is relatively strong but the two heritability estimates are roughly the same. (For detailed analysis and examples, see [SI Appendix, section D](#).) Given these two

issues, the dichotomy of nature and nurture simply does not apply in the systems with genetic nurturing and vertical cultural transmission.

### Population Subdivision and Assortative Mating

Population subdivision and assortative mating can have profound effects on the missing heritability (24, 25). It is then interesting to compare statistical signals generated by population subdivision and assortative mating with that of genetic nurturing. Kong et al. (26) made a correction for assortative mating. Here we use a simple one-locus, two-allele purely genetic model to show how population subdivision and assortative mating can generate a signal similar to that of genetic nurturing, namely, the correlation between parents' nontransmitted alleles and children's phenotypes. We also examine how Kong et al. (26) corrected for assortative mating.

**The Effect of Population Subdivision.** Consider a metapopulation  $H$  which includes  $k$  subpopulations, denoted by  $H_1, \dots, H_k$ . The relative sizes of the subpopulations are denoted by  $\alpha_1, \dots, \alpha_k$ , respectively. Again we assume there are two alleles,  $A_1$  and  $A_2$ , and two phenotypes, bar and nonbar. The probability of being bar is genetically determined and is  $c_1$  for  $A_1A_1$ ,  $c_2$  for  $A_1A_2$ , and  $c_3$  for  $A_2A_2$ . Hardy–Weinberg equilibrium is assumed for each subpopulation and the frequencies of  $A_1$  and  $A_2$  in population  $i$  are assumed to be  $p_i$  and  $q_i$ . The frequencies of  $A_1$  and  $A_2$  in the metapopulation are thus  $p = \alpha_1 p_1 + \dots + \alpha_k p_k$  and  $q = \alpha_1 q_1 + \dots + \alpha_k q_k$ , respectively. The genotype frequencies for  $A_1A_1, A_1A_2, A_2A_2$  are then  $\sum_{i=1}^k \alpha_i p_i^2, 2 \sum_{i=1}^k \alpha_i p_i q_i, \sum_{i=1}^k \alpha_i q_i^2$ , respectively, which we denote by  $u, 2v, w$ . Set  $x_i = c_1 p_i^2 + 2c_2 p_i q_i + c_3 q_i^2$  and  $x = \sum_{i=1}^k \alpha_i x_i$ , which represent, respectively, the frequencies of bar in each subpopulation and the metapopulation. We now study how subpopulation structure may mimic genetic nurturing.

First, we regress the child's phenotype on alleles transmitted and alleles not transmitted to the child (in the metapopulation). We again denote the child's phenotypic value by  $X_{ch}$ , which is 1 for the bar phenotype and 0 for the nonbar phenotype. Let the numbers of the father's and mother's transmitted  $A_2$  alleles be  $(N_p^t, N_m^t)$  and the numbers of the nontransmitted  $A_2$  alleles be  $(N_p^n, N_m^n)$ . The regression is then

$$X_{ch} = \beta_1(N_p^t + N_m^t) + \beta_2(N_p^n + N_m^n) + \delta, \quad [13]$$

where  $\beta_1$  and  $\beta_2$  are the regression coefficients and  $\delta$  is the sum of the intercept and the residual.

Using the law of total variance we can calculate the covariances among  $N_p^t, N_m^t, N_p^n, N_m^n$  and  $X_{ch}$  (SI Appendix, section F) and obtain

$$\begin{aligned} Cov(N_A^a, N_B^b) &= pq - v; \\ Cov(X_{ch}, N_m^n) &= Cov(X_{ch}, N_p^n) = \sum_{i=1}^k \alpha_i q_i x_i - qx; \\ Cov(X_{ch}, N_m^t) &= Cov(X_{ch}, N_p^t) = \sum_{i=1}^k \alpha_i q_i x_i - qx \\ &\quad + \sum_{i=1}^k \alpha_i p_i q_i [p_i(c_2 - c_1) + q_i(c_3 - c_2)], \end{aligned} \quad [14]$$

where  $N_A^a$  and  $N_B^b$  are two different elements of the set  $\{N_p^t, N_m^t, N_p^n, N_m^n\}$ . Denote  $\sum_{i=1}^k \alpha_i q_i x_i - qx$  by  $K_1$  and  $\sum_{i=1}^k \alpha_i p_i q_i [p_i(c_2 - c_1) + q_i(c_3 - c_2)]$  by  $K_2$ . Then from multivariate regression analysis (SI Appendix, section F), we obtain

$$\begin{aligned} \beta_1 &= \frac{K_1 v + K_2(2pq - v)}{(4pq - 3v)v}; \\ \beta_2 &= \frac{K_1 v - K_2(2pq - 2v)}{(4pq - 3v)v}; \end{aligned} \quad [15]$$

and

$$\beta_1 - \beta_2 = \frac{K_2}{v}. \quad [16]$$

Thus, when a population is subdivided, the nontransmitted alleles will also be correlated with the phenotype and the regression coefficient  $\beta_2$  will in general not be zero. This correlation is generated by differences among allele frequencies in subpopulations. Both  $K_1$ , the covariance of probabilities being bar among subpopulations, and  $pq - v$ , the variance of allele frequencies among subpopulations, will be nonzero when such differences exist. This means that even in the purely genetic case, population subdivision may generate correlation between children's phenotypes and nontransmitted alleles, which is considered by Kong et al. (26) as the signal of genetic nurturing. Thus, a correction must be made if we want to separate the genetic nurturing effect from that of population subdivision.

**The Effect of Assortative Mating.** Assortative mating can be viewed as a special kind of population subdivision, namely, division into assorting groups, and an analysis similar to the above can be applied with assortative mating. Again we have alleles  $A_1$  and  $A_2$  with phenotypes bar and nonbar. The frequencies of  $A_1$  and  $A_2$  are  $p$  and  $q$ . The probabilities of a child being bar if its genotype is  $A_1A_1, A_1A_2, A_2A_2$  are, respectively,  $c_1, c_2, c_3$ . In each generation, we assume a proportion  $r$  of individuals mate assortatively based on phenotype and the remaining individuals mate randomly. Denote the genotype frequencies of  $A_1A_1, A_1A_2, A_2A_2$  in generation  $n$  by  $P^n, 2Q^n, R^n$ , respectively. Since the proportion of bar in each genotype is fixed, we need only analyze the genotype frequency dynamics, and since  $P^n + Q^n = p; R^n + Q^n = q$  and  $p$  and  $q$  do not change, there is only one free variable. Thus, we need only to solve the recursion for one of  $P^n, Q^n, R^n$ ; here we focus on the recursion for  $Q^n$ , namely

$$\begin{aligned} Q^{n+1} &= (1-r)pq + \left\{ \frac{(P^n c_1 + Q^n c_2)(R^n c_3 + Q^n c_2)}{[P^n c_1 + 2Q^n c_2 + R^n c_3]^2} \times \right. \\ &\quad \left. (rP^n c_1 + 2rQ^n c_2 + rR^n c_3) \right\} \\ &\quad + \left\{ \frac{[P^n(1-c_1) + Q^n(1-c_2)][R^n(1-c_3) + Q^n(1-c_2)]}{[P^n(1-c_1) + 2Q^n(1-c_2) + R^n(1-c_3)]^2} \right. \\ &\quad \left. \times [rP^n(1-c_1)] \right\} \\ &\quad + 2rQ^n(1-c_2) + rR^n(1-c_3). \end{aligned} \quad [17]$$

In SI Appendix, section G, we prove that there is a unique globally stable fixed point for the dynamical system 17. Assuming the equilibrium genotype frequencies are  $(u, 2v, w)$ , we can use a regression analysis similar to that used for population subdivision. Let

$$\begin{aligned} \alpha &= [p(c_2 - c_3) - q(c_2 - c_1)]v + pq(c_3 - c_1); \\ \beta &= (c_3 + c_1 - 2c_2)(q_1 + q_2) + 2(c_2 - c_1); \\ \bar{x} &= E(X_{ch}) = uc_1 + 2vc_2 + wc_3. \end{aligned} \quad [18]$$

We can then compute the covariances among  $N_p^t, N_m^t, N_p^n, N_m^n$  and  $X_{ch}$  (SI Appendix, section H); namely,



$$\begin{aligned}
Cov(N_A^a, N_B^b) &= \frac{r\alpha^2}{\bar{x}(1-\bar{x})}; \\
Cov(X_{ch}, N_m^n) &= Cov(X_{ch}, N_p^n) = \alpha; \\
Cov(X_{ch}, N_m^t) &= Cov(X_{ch}, N_p^t) = \beta Cov(N_A^a, N_B^b) = \frac{r\alpha^2\beta}{\bar{x}(1-\bar{x})}.
\end{aligned}
\tag{19}$$

Comparing these results with those of the population subdivision model, we see that in the assortative mating model,

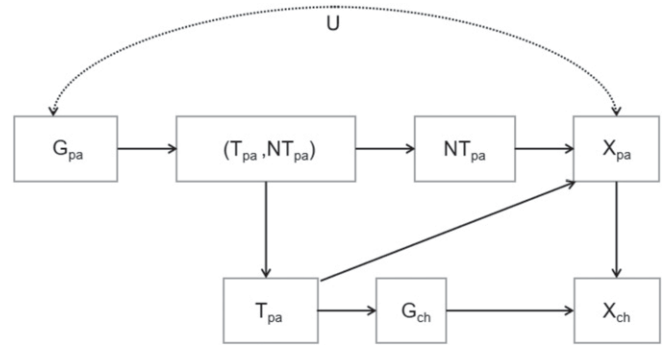
$$\begin{aligned}
K_1 &= Cov(X_{ch}, N_m^n) = \beta(pq - v); \\
K_2 &= Cov(X_{ch}, N_m^t) - Cov(X_{ch}, N_p^n) = \alpha\left[1 - \frac{r\alpha\beta}{\bar{x}(1-\bar{x})}\right],
\end{aligned}$$

which, together with Eq. 16, gives the regression coefficients. As a special case of population subdivision, assortative mating can also generate a similar signal to that of genetic nurturing and should be corrected for in computations of correlations. Kong et al. (26) do make the correction and we agree with their conclusion that the effect of assortative mating is relatively weak. However, there may be a problem with their model assumptions. To be specific, when assumption VI in their supplementary material section “Estimating the confounding effects induced by assortative mating” is translated into our model, it amounts to assuming  $Cov(N_p^t, N_p^n)$  and  $Cov(N_m^t, N_m^n)$  are zero while  $Cov(N_p^t, N_m^t)$ ,  $Cov(N_p^t, N_m^n)$ ,  $Cov(N_p^t, N_m^t)$ ,  $Cov(N_m^t, N_p^n)$ ,  $Cov(N_p^t, N_m^t)$ , and  $Cov(N_p^t, N_m^n)$  are not zero. From our analysis this appears unreasonable since these covariances should all be the same.

### A Unified Causal Framework for Genetic Nurturing, Population Subdivision, and Assortative Mating

In previous sections, we proposed a mechanistic model for genetic nurturing and showed how population subdivision and assortative mating could mimic genetic nurturing. We followed Kong et al. (26) and used a linear regression (i.e., path analysis or structural equation model [SEM]) method for mediation analysis. We showed that this method works well in the linear setting but becomes problematic in the nonlinear case. Fortunately, traditional structural mediation analysis and separation of total effect into direct effect and indirect effect can be naturally extended to the nonlinear setting using *do* calculus and counterfactuals in causal analysis (38). Further, causal analysis provides powerful techniques for analyzing different causal effects and the influence of confounding factors (39, 40), which can also help frame the mismatch between causality and statistics used in the nature–nurture discussion, such as variance (6) and correlation. We therefore employ causal analysis in constructing a semi-Markovian causal model (Fig. 1) for genetic nurturing, population subdivision, and assortative mating (details in *SI Appendix, section J*). Using this unified causal framework, we show that the correlations between non-transmitted alleles and the phenotype under genetic nurturing, population subdivision, and assortative mating represent different pathways in the causal diagram. We also analyze different effects of transmitted alleles and clarify when the analysis by Kong et al. (26) of genetic nurturing works and when it fails.

In Fig. 1,  $G_{pa} = (G_m, G_p)$  and  $G_{ch}$ , respectively, represent the parents’ and the child’s genotypes,  $T_{pa} = (T_m, T_p)$  represents the alleles transmitted from the parents, and  $NT_{pa} = (NT_m, NT_p)$  represents the alleles not transmitted from the parents.  $X_{pa} = (X_m, X_p)$  and  $X_{ch}$  represent the parents’ and the child’s phenotypes. The parents’ genotypes and phenotypes are confounded by  $U$  and are thus connected by the bidirectional arc in Fig. 1. We study the effects of transmitted alleles and nontransmitted alleles.



**Fig. 1.** A unified causal diagram for genetic nurturing, population subdivision, and assortative mating.

In the causal diagram of Fig. 1, the “total effect” of changing transmitted alleles from  $T_{pa} = x$  to  $T_{pa} = y$  on the child’s phenotype can be represented by  $P(X_{ch}|do(T_{pa} = y)) - P(X_{ch}|do(T_{pa} = x))$ , in which case the quantity corresponding to the regression coefficient  $\beta_1$  will be  $P(X_{ch}|T_{pa} = y) - P(X_{ch}|T_{pa} = x)$ ; the total effect of changing nontransmitted alleles from  $NT_{pa} = x$  to  $NT_{pa} = y$  on the child’s phenotype can be represented by  $P(X_{ch}|do(NT_{pa} = y)) - P(X_{ch}|do(NT_{pa} = x))$ , in which case the quantity corresponding to the regression coefficient  $\beta_2$  will be  $P(X_{ch}|NT_{pa} = y) - P(X_{ch}|NT_{pa} = x)$ . The relationships among these quantities and the path-specific effects corresponding to the “direct effect” and the “indirect effect” in Kong et al.’s (26) paper are analyzed in detail below.

Before doing the analysis, we address a key problem for the identification of different effects in Fig. 1, namely, the definition of  $P(X_{pa}|T_{pa}, NT_{pa}, G_{pa})$ . This quantity is not always well defined because the parents’ genotypes are automatically determined when the transmitted alleles and nontransmitted alleles are known. Thus, to use causal analysis, we must define this probability  $P(X_{pa}|T_{pa}, NT_{pa}, G_{pa})$  when  $G_{pa}$  is not compatible with  $T_{pa}, NT_{pa}$ . If the confounding factor  $U$  in Fig. 1 is known, then this quantity is just  $\sum_U P(X_{pa}|T_{pa}, NT_{pa}, U)P(U|G_{pa})$ ; otherwise, some prior knowledge of  $U$  is needed. Table 2 lists the symbols used throughout our application of the causal analysis, while Table 3 lists the analogous symbols used in the cultural transmission model.

Fortunately, for genetic nurturing, population subdivision, and assortative mating, the required prior knowledge exists. For population subdivision and assortative mating, since the effect is purely genetic, we have  $P(X_{pa}|T_{pa}, NT_{pa}, G_{pa}) = P(X_{pa}|T_{pa}, NT_{pa})$ . For genetic nurturing,  $U$  represents the grandparents’ phenotypic influence. Denote the mother’s parents’ phenotypes by  $X_{mpa} = (X_{mm}, X_{mp})$  and the father’s parents’ phenotypes by  $X_{ppa} = (X_{pm}, X_{pp})$ . Then

$$\begin{aligned}
&P(X_{pa}|T_{pa}, NT_{pa}, G_{pa}) \\
&= \sum_{\substack{X_{mpa}, \\ X_{ppa}}} P(X_{pa}|T_{pa}, NT_{pa}, X_{mpa}, X_{ppa})P(X_{mpa}, X_{ppa}|G_{pa}),
\end{aligned}
\tag{20}$$

which can be calculated if data from the grandparents’ generation are given. If these data are not given (which we assume in this paper), we may assume that the population is at equilibrium under random mating, in which case

$$\begin{aligned}
&P(X_{pa}|T_{pa}, NT_{pa}, X_{mpa}, X_{ppa}) \\
&= P(X_m|T_m, NT_m, X_{mpa})P(X_p|T_p, NT_p, X_{ppa}); \\
&P(X_{mpa}, X_{ppa}|G_{pa}) = P(X_{mpa}|G_m)P(X_{ppa}|G_p).
\end{aligned}
\tag{21}$$

**Table 2. Symbols used in causal analysis**

Notation	Meaning
$G_{ch}$	The child's genotype
$G_m$	The mother's genotype
$G_p$	The father's genotype
$G_{pa}$	The parents' genotypes, denoted by $(G_m, G_p)$
$X_{ch}$	The child's phenotype
$X_m$	The mother's phenotype
$X_p$	The father's phenotype
$X_{pa}$	The parents' phenotypes, denoted by $(X_m, X_p)$
$X_{mm}$	The mother's mother's phenotype
$X_{mp}$	The mother's father's phenotype
$X_{mpa}$	The mother's parents' phenotypes, denoted by $(X_{mm}, X_{mp})$
$X_{pm}$	The father's mother's phenotype
$X_{pp}$	The father's father's phenotype
$X_{ppa}$	The father's parents' phenotypes, denoted by $(X_{pm}, X_{pp})$
$T_m$	Alleles transmitted from the mother to the child
$T_p$	Alleles transmitted from the father to the child
$T_{pa}$	Alleles transmitted from the parents to the child, denoted by $(T_m, T_p)$
$NT_m$	Alleles not transmitted from the mother to the child
$NT_p$	Alleles not transmitted from the father to the child
$NT_{pa}$	Alleles not transmitted from the parents to the child, denoted by $(NT_m, NT_p)$

Each term on the right-hand side of [21] can be calculated using data from the parents' generation. For example, we may assume  $P(X_{mpa} = x | G_m = g) = P(X_{pa} = x | G_{ch} = g)$ , where  $x$  is a vector of two phenotypes and  $g$  is a genotype. ( $P(X_m | T_m, NT_m, G_m)$  and  $P(X_p | T_p, NT_p, G_p)$  can be defined similarly.)

Now, since  $P(X_{pa} | T_{pa}, NT_{pa}, G_{pa})$  is defined and identifiable, we are now able to calculate the effects of nontransmitted alleles ( $NT_{pa}$ ) on the child's phenotype ( $X_{ch}$ ).  $NT_{pa}$  has one front-door (causal) path (we denote this by path 1; for definition of front-door path, see *SI Appendix, section J*) to  $X_{ch}$ ,

$$NT_{pa} \rightarrow X_{pa} \rightarrow X_{ch},$$

and three backdoor (noncausal) paths (denoted by paths 2, 3, and 4; for definition of backdoor path, see *SI Appendix, section J*) to  $X_{ch}$ ,

$$\begin{aligned} NT_{pa} &\leftarrow (T_{pa}, NT_{pa}) \rightarrow T_{pa} \rightarrow G_{ch} \rightarrow X_{ch}; \\ NT_{pa} &\leftarrow (T_{pa}, NT_{pa}) \rightarrow T_{pa} \rightarrow X_{pa} \rightarrow X_{ch}; \\ NT_{pa} &\leftarrow (T_{pa}, NT_{pa}) \leftarrow G_{pa} \leftrightarrow X_{pa} \rightarrow X_{ch}. \end{aligned}$$

Following refs. 40–42, in the causal diagram Fig. 1,  $P(X_{ch} | do(NT_{pa}))$  can be given by

$$P(X_{ch} | do(NT_{pa})) = \sum_{\substack{G_{pa} \\ T_{pa} \\ X_{pa}}} \{P(G_{pa})P(T_{pa} | G_{pa}) \times P(X_{pa} | T_{pa}, NT_{pa}, G_{pa})P(X_{ch} | T_{pa}, X_{pa})\}. \quad [22]$$

Thus the total effect of changing from  $NT_{pa} = x$  to  $NT_{pa} = y$  given by  $P(X_{ch} | do(NT_{pa} = y)) - P(X_{ch} | do(NT_{pa} = x))$  can be calculated using Eq. 22. In the population subdivision and assortative mating case, since there is no effect of the parental phenotypes, the edge  $X_{pa} \rightarrow X_{ch}$  (in Fig. 1) and paths 1, 3, and 4 vanish, so  $P(X_{ch} | G_{ch}, X_{pa}) = P(X_{ch} | G_{ch})$ . This automatically gives  $P(X_{ch} | do(NT_{pa})) = P(X_{ch})$  and  $P(X_{ch} | do(NT_{pa} = y)) - P(X_{ch} | do(NT_{pa} = x)) = 0$ , which means there is no causal effect of the nontransmitted alleles on the child's phenotype. However, because of the existence of

the active backdoor path 2,  $P(X_{ch} | NT_{pa}) \neq P(X_{ch})$  and  $P(X_{ch} | NT_m = y) - P(X_{ch} | NT_m = x)$  is not zero in general. To be specific, population subdivision and assortative mating introduce correlation between allele frequencies and (assorting) groups so that  $NT_{pa}$  contains group information. As a result,  $P(T_{pa} | NT_{pa} = y)$  and  $P(T_{pa} | NT_{pa} = x)$  will generally not be the same and  $P(X_{ch} | NT_{pa} = y) - P(X_{ch} | NT_{pa} = x) = \sum_{T_{pa}} [P(T_{pa} | NT_{pa} = y) - P(T_{pa} | NT_{pa} = x)] P(X_{ch} | T_{pa})$  will not be zero in general, which explains the correlation between the nontransmitted alleles and the child's phenotype when there is population subdivision or assortative mating despite there being no causal path between these two quantities.

For genetic nurturing,  $G_{pa}$  and  $X_{pa}$  are confounded due to the influence of the grandparents' phenotypes. In a random mating population, paths 2 and 3 are deactivated because under random mating,  $P(T_{pa} | NT_{pa}) = P(T_{pa} | do(NT_{pa})) = P(T_{pa})$ . Thus,  $P(X_{ch} | do(NT_{pa}))$  is influenced by paths 1 and 4, with  $P(X_{ch} | do(NT_{pa}))$  representing the effect of path 1. This means that in the genetic nurturing case,  $P(X_{ch} | NT_{pa} = y) - P(X_{ch} | NT_{pa} = x)$ , which corresponds to  $\beta_2$  in the previous section (Eqs. 6 and 8b), is in general not a suitable representation of the effect of the nontransmitted alleles. The correlation between the nontransmitted alleles and the child's phenotype is in general not zero, but is a result of both genetic nurturing and confounding between the parents' genotypes and phenotypes. This confounding, in the genetic nurturing case, is due to the grandparents' phenotypic influence. [Kong et al. (26) did not take account of the grandparents' influence or genetic nurturing's influence on pheno-genotype frequencies; they therefore artificially deactivated path 4, which allowed their regression to work.]

Now we consider the effect of transmitted alleles. It is easy to see that there are four paths from  $T_{pa}$  to  $X_{ch}$ , namely, two front-door paths (denoted by path 1' and path 2')

$$\begin{aligned} T_{pa} &\rightarrow G_{ch} \rightarrow X_{ch}; \\ T_{pa} &\rightarrow X_{pa} \rightarrow X_{ch}, \end{aligned}$$

and two backdoor paths (denoted by path 3' and path 4')

$$\begin{aligned} T_{pa} &\leftarrow (T_{pa}, NT_{pa}) \rightarrow NT_{pa} \rightarrow X_{pa} \rightarrow X_{ch}; \\ T_{pa} &\leftarrow (T_{pa}, NT_{pa}) \leftarrow G_{pa} \leftrightarrow X_{pa} \rightarrow X_{ch}. \end{aligned}$$

From the analyses in refs. 40–42,  $P(X_{ch} | do(T_{pa}))$  is given by

$$P(X_{ch} | do(T_{pa})) = \sum_{\substack{G_{pa} \\ NT_{pa} \\ X_{pa}}} \{P(G_{pa})P(NT_{pa} | G_{pa}) \times P(X_{pa} | T_{pa}, NT_{pa}, G_{pa})P(X_{ch} | T_{pa}, X_{pa})\} \quad [23]$$

**Table 3. Matching symbols in the two models**

Notation in cultural transmission model	Corresponding representation in the unified causal model
$c_1$	$P(X_{ch} = 1   G_{ch} = A_1 A_1, X_{pa} = (0, 0))$
$c_1 + e_1$	$P(X_{ch} = 1   G_{ch} = A_1 A_1, X_{pa} = (1, 0) \text{ or } (0, 1))$
$c_1 + 2e_1$	$P(X_{ch} = 1   G_{ch} = A_1 A_1, X_{pa} = (1, 1))$
$c_2$	$P(X_{ch} = 1   G_{ch} = A_1 A_2, X_{pa} = (0, 0))$
$c_2 + e_2$	$P(X_{ch} = 1   G_{ch} = A_1 A_2, X_{pa} = (1, 0) \text{ or } (0, 1))$
$c_2 + 2e_2$	$P(X_{ch} = 1   G_{ch} = A_1 A_2, X_{pa} = (1, 1))$
$c_3$	$P(X_{ch} = 1   G_{ch} = A_2 A_2, X_{pa} = (0, 0))$
$c_3 + e_3$	$P(X_{ch} = 1   G_{ch} = A_2 A_2, X_{pa} = (1, 0) \text{ or } (0, 1))$
$c_3 + 2e_3$	$P(X_{ch} = 1   G_{ch} = A_2 A_2, X_{pa} = (1, 1))$

Here for simplicity, we assume  $X = 1$  for bar phenotype and  $X = 0$  for nonbar phenotype.

Similarly as in the above analysis of nontransmitted alleles, with population subdivision or assortative mating, only path 1' is activated, and  $P(X_{ch}|do(T_{pa})) = P(X_{ch}|T_{pa})$ . With genetic nurturing, path 3' vanishes while paths 1', 2', and 4', which jointly contribute to  $P(X_{ch}|T_{pa})$ , are activated. Also  $P(X_{ch}|T_{pa})$  represents the joint effect of causal paths 1' and 2'. As for the nontransmitted alleles, because of the confounding of the parents' genotypes and phenotypes, which introduces a back-door path 4',  $P(X_{ch}|T_{pa} = y) - P(X_{ch}|T_{pa} = x)$ , corresponding to  $\beta_1$  (Eqs. 6 and 8a), is in general not a suitable representation of the total effect of the transmitted alleles. [Again Kong et al. (26) ignore path 4', which allows their regression method to work.]

Nonetheless, the total effect of the transmitted alleles can be calculated using the *do* operator. Since there are two causal paths in this case, it is interesting to identify the path-specific effects. In mediation analysis, the path-specific effects of paths 1' and 4' are called "natural direct effect" and "natural indirect effect," respectively. The total effect of changing from  $T_{pa} = x$  to  $T_{pa} = y$  can then be partitioned to into these two effects. Following refs. 43 and 44, the natural direct and natural indirect effects can be calculated as follows:

Natural direct effect:

$$\sum_{\substack{G_{pa} \\ NT_{pa} \\ X_{pa}}} \{P(G_{pa})P(NT_{pa}|G_{pa})P(X_{pa}|T_{pa} = x, NT_{pa}, G_{pa}) \\ \times [P(X_{ch}|T_{pa} = y, X_{pa}) - P(X_{ch}|T_{pa} = x, X_{pa})]\}; \quad [24]$$

Natural indirect effect:

$$\sum_{\substack{G_{pa} \\ NT_{pa} \\ X_{pa}}} \{P(G_{pa})P(NT_{pa}|G_{pa})P(X_{ch}|T_{pa} = y, X_{pa}) \times \\ [P(X_{pa}|T_{pa} = y, NT_{pa}, G_{pa}) - P(X_{pa}|T_{pa} = x, NT_{pa}, G_{pa})]\}. \quad [25]$$

Kong et al. (26) also make this separation and called them direct effect and indirect effect, respectively. They also use the difference between regression coefficients for the transmitted alleles and the nontransmitted alleles to represent the direct effect of transmitted alleles. However, as we have shown above, their estimates of the total effects of the transmitted and nontransmitted alleles, represented by  $P(X_{ch}|T_{pa} = y) - P(X_{ch}|T_{pa} = x)$  and  $P(X_{ch}|NT_{pa} = y) - P(X_{ch}|NT_{pa} = x)$ , respectively, are biased due to the confounding of the parents' phenotypes and genotypes. Thus, we need to analyze the relationships among the following three quantities:

- 1) The difference between the total effects of the transmitted alleles and the nontransmitted alleles, represented by

$$[P(X_{ch}|do(T_{pa} = y)) - P(X_{ch}|do(T_{pa} = x))] \\ - [P(X_{ch}|do(NT_{pa} = y)) - P(X_{ch}|do(NT_{pa} = x))]. \quad [26]$$

- 2) Kong et al.'s (26) estimates of direct effect, represented by

$$[P(X_{ch}|T_{pa} = y) - P(X_{ch}|T_{pa} = x)] \\ - [P(X_{ch}|NT_{pa} = y) - P(X_{ch}|NT_{pa} = x)]. \quad [27]$$

- 3) Path 1-specific causal effects, the natural direct effect in our causal diagram (Fig. 1), represented by

$$\sum_{\substack{G_{pa}, \\ NT_{pa}, \\ X_{pa}}} \{P(G_{pa})P(NT_{pa}|G_{pa})P(X_{pa}|T_{pa} = x, NT_{pa}, G_{pa}) \\ \times [P(X_{ch}|T_{pa} = y, X_{pa}) - P(X_{ch}|T_{pa} = x, X_{pa})]\}. \quad [28]$$

These three quantities will be the same (for proof, see *SI Appendix, section J*) if there exist  $F_1, F_{1'}, F_2$  such that for  $y = (y_m, y_p)$  and  $x = (x_m, x_p)$

$$P(X_m|T_m, NT_m = y_m, G_m) \\ - P(X_m|T_m, NT_m = x_m, G_m) = F_1(X_m, y_m, x_m); \quad [29]$$

$$P(X_m|T_m = y_m, NT_m, G_m) \\ - P(X_m|T_m = x_m, NT_m, G_m) = F_1(X_m, y_m, x_m); \quad [30]$$

$$P(X_p|T_p, NT_p = y_p, G_p) \\ - P(X_p|T_p, NT_p = x_p, G_p) = F_{1'}(X_p, y_p, x_p); \quad [31]$$

$$P(X_p|T_p = y_p, NT_p, G_p) \\ - P(X_p|T_p = x_p, NT_p, G_p) = F_{1'}(X_p, y_p, x_p); \quad [32]$$

$$P(X_{ch}|T_{pa} = y, X_{pa}) \\ - P(X_{ch}|T_{pa} = x, X_{pa}) = F_2(X_{ch}, y, x). \quad [33]$$

Obviously, this assumption of additivity is exactly what the SEM and that of Kong et al. (26) assume. However, there are two problems. First, even when this assumption holds and the direct effect is estimated properly, their estimate of the total effect and the indirect effect is biased (*SI Appendix, section J*). Second, when this assumption is violated, as in the general case of the model defined by Table 1, these three quantities will not be the same. Thus, using the difference between total effects of the transmitted and nontransmitted alleles to estimate the direct effect of the transmitted alleles may be problematic.

## Discussion

In this paper, we propose a simple generative model of cultural transmission and study the effects of including population subdivision and assortative mating in the model. Four main issues are addressed. First, the cultural transmission model gives a quantitative explanation of genetic nurturing that can be compared with that of Kong et al. (26). Second, we show how the notion of heritability from traditional statistical genetics is inadequate under this simple setting. Third, we show how population subdivision and assortative mating can generate signals similar to genetic nurturing and examine Kong et al.'s method (26) for correction of assortative mating. Finally, we propose a unified causal diagram for genetic nurturing, population subdivision, and assortative mating. Two statistical signals are used for our analysis: the difference between heritability estimated from GWAS and that estimated from pedigrees and the nonnegligible correlation between parents' nontransmitted alleles and children's phenotypes. Here we discuss the meaning of these two signals in a more general context.

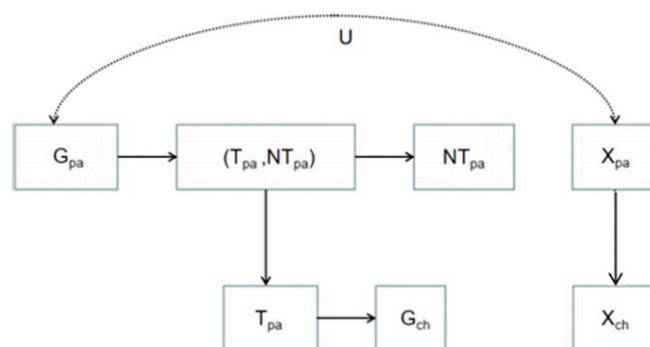
For the first signal, represented by the missing heritability, we show in our gene-culture coevolution model, that heritability is an ill-defined notion and fails to decompose the total variance into nature and nurture. In addition, this failure, which is reflected in the missing heritability problem, has to be considered in the broader context of other problems, since it can be caused by both genetic and demographic factors, whose importance can vary from trait to trait. For example, having too few SNPs may produce missing heritability in height and body mass index (21, 22), while for social and behavioral traits, cultural transmission can be an important factor. However, even for height and body-mass index (BMI), it is not easy to exclude confounding

cultural factors, such as diet, that can be vertically transmitted. In our analysis, when genetic nurturing exists, we find parameter regions where cultural transmission is not negligible, but the two estimates of heritability are roughly the same; this can be easily extended to the cases where parents' phenotypes and genotypes are confounded by factors other than genetic nurturing. Thus, when vertical cultural transmission is potentially involved, heritability will not be a meaningful statistic.

Second, we show that genetic nurturing, population subdivision, and assortative mating can contribute to a second signal, namely, the correlation between parents' nontransmitted alleles and children's phenotypes. We also show that the method Kong et al. (26) used to correct for assortative mating entails restrictive assumptions. Using the causal diagram in Fig. 1, we see that the scenario is more complex when genetic nurturing coexists with population subdivision or assortative mating. When genetic nurturing coexists with population subdivision,  $P(X_{pa}|T_{pa}, NT_{pa}, G_{pa})$  can be defined similarly for each subpopulation. The difference between this case and genetic nurturing without population subdivision is that paths 2, 3, and 3' will be activated. With genetic nurturing in the case of assortative mating, there is an added complication due to the additional source of confounding between the parents' genotypes and phenotypes; prior information about this confounding is needed for  $P(X_{pa}|T_{pa}, NT_{pa}, G_{pa})$  to be well defined.

In our unified causal diagram (Fig. 1), the correlation between the parents' nontransmitted alleles ( $NT_{pa}$ ) and the child's phenotype ( $X_{ch}$ ) can arise from two different classes of paths: either through  $T_{pa}$  (paths 2 and 3) or not through  $T_{pa}$  but through  $X_{pa}$  (paths 1 and 4). The paths through  $T_{pa}$  can be tested and possibly corrected using linkage disequilibrium (LD). The path through  $X_{pa}$ , either causal or noncausal, implies the influence of parents' phenotypes. Here we emphasize two points. First, although in the two-allele, two-phenotype model, we assumed that the parents' phenotypes and that of the child referred to the same trait, in our causal diagram they could refer to different traits that have a causal relationship. In this case there will be "cultural influence" instead of "cultural transmission" but the whole causal diagram and all of the analysis will still be valid. Second, although without prior information about confounding the different effects cannot be calculated, confounding factors other than genetic nurturing between the parents' phenotypes and genotypes can still produce correlation between the nontransmitted alleles and the child's phenotype if the pathway from  $X_{pa}$  to  $X_{ch}$  is activated, even when there is no genetic basis for the phenotypes. Path 4 still exists even if  $NT_{pa} \rightarrow X_{pa}$ ,  $T_{pa} \rightarrow X_{pa}$ , and  $G_{ch} \rightarrow X_{ch}$  do not exist, see Fig. 2. (In principle, we should also include  $G_{ch} \leftrightarrow X_{ch}$  in Fig. 2 since the genotypes and phenotypes are confounded in the parents' generation. However, adding this will not activate any path, so we neglect it for simplicity.) In this case the correlation generated by the parental phenotypic influence and genotype-phenotype confounding will be an indicator of pure cultural influence or cultural transmission.

Although our causal diagram unifies genetic nurturing, population subdivision, and assortative mating and can be extended qualitatively to more general cases of linkage disequilibrium and parental influences, it does not exhaust all potential mechanisms that could generate correlation between nontransmitted



**Fig. 2.** A potential causal diagram where  $NT_{pa}$  and  $X_{ch}$  are correlated due to pure cultural transmission.

alleles and children's phenotypes. Recent findings by Mostafavi et al. (45) that polygenic scores, which are computed under additive assumptions on allelic variances, are not portable among population groups with the same apparent genetic ancestry are relevant in the context of our analysis. They attribute their result to factors such as genetic nurturing (i.e., indirect genetic affects) and assortative mating studied here, as well as varying levels of environmental variance and genotype-by-environment interactions.

The correlation between parents' nontransmitted alleles and children's phenotypes is a signal of demographic or social influences, which can include parental phenotypic influence (genetic nurturing as a special case), or linkage disequilibrium (population subdivision as a special case). (Confounding between  $G_{pa}$  and  $X_{ch}$  without referring to  $X_{pa}$  can also contribute to such correlation, but in reality will be less likely.) More knowledge of the potential confounding factors is needed if we want to distinguish between these mechanisms more precisely.

Lewontin (ref. 6, p. 409) pointed out the pitfalls in inferring causality from variance statistics, including heritability derived from familial analyses: "The analysis of causes in human genetics is meant to provide us with the basic knowledge we require for correct schemes of environmental modification and intervention. Together with a knowledge of the relative frequencies of different human genotypes, a knowledge of norms of reaction can also predict the demographic and public health consequences of certain massive environmental changes. Analysis of variance can do neither of these because its results are a unique function of the present distribution of environment and genotypes." Lewontin's admonition, made 46 y ago, remains pertinent today with respect to SNP heritability derived from GWAS. Inference of causality from the relationships among statistical signals, including correlations and variances, remains extremely difficult.

**Data Availability.** All study data are included in this article and [SI Appendix](#).

**ACKNOWLEDGMENTS.** This research was supported in part by the Morrison Institute for Population and Resource Studies at Stanford University and by the Stanford Center for Computational, Evolutionary and Human Genomics. We thank Professors Jonathan Pritchard and Ewart Thomas for their comments on an earlier draft.

1. J. L. Lush, *Animal Breeding Plans* (Iowa State College Press, ed. 2, 1943).
2. A. Jensen, How much can we boost IQ and scholastic achievement? *Harv. Educ. Rev.* **39**, 1–123 (1969).
3. T. J. Bouchard, M. McGue, Familial studies of intelligence: A review. *Science* **212**, 1055–1059 (1981).
4. R. J. Herrnstein, C. Murray, *The Bell Curve: Intelligence and Class Structure in American Life* (Simon and Schuster, 2010).
5. L. J. Kamin, *The Science and Politics of IQ* (Routledge, 2012).
6. R. C. Lewontin, Annotation: The analysis of variance and the analysis of causes. *Am. J. Hum. Genet.* **26**, 400–411 (1974).

7. M. W. Feldman, R. C. Lewontin, The heritability hang-up. *Science* **190**, 1163–1168 (1975).
8. M. W. Feldman, F. B. Christiansen, S. P. Otto, Gene-culture co-evolution: Teaching, learning, and correlations between relatives. *Isr. J. Ecol. Evol.* **59**, 72–91 (2013).
9. E. Turkheimer, Weak genetic explanation 20 years later: Reply to Plomin et al. (2016). *Perspect. Psychol. Sci.* **11**, 24–28 (2016).
10. L. L. Cavalli-Sforza, M. W. Feldman, Cultural versus biological inheritance: Phenotypic transmission from parents to children. (A theory of the effect of parental phenotypes on children's phenotypes). *Am. J. Hum. Genet.* **25**, 618–637 (1973).



11. D. C. Rao, N. E. Morton, S. Yee, Analysis of family resemblance. II. A linear model for familial correlation. *Am. J. Hum. Genet.* **26**, 331–359 (1974).
12. D. C. Rao, N. E. Morton, S. Yee, Resolution of cultural and biological inheritance by path analysis. *Am. J. Hum. Genet.* **28**, 228–242 (1976).
13. C. R. Cloninger, J. Rice, T. Reich, Multifactorial inheritance with cultural transmission and assortative mating. I. Description and basic properties of the unitary models. *Am. J. Hum. Genet.* **30**, 618–643 (1978).
14. C. R. Cloninger, J. Rice, T. Reich, Multifactorial inheritance with cultural transmission and assortative mating. II. A general model of combined polygenic and cultural inheritance. *Am. J. Hum. Genet.* **31**, 176–198 (1979).
15. C. R. Cloninger, J. Rice, T. Reich, Multifactorial inheritance with cultural transmission and assortative mating. III. Family structure and the analysis of separation experiments. *Am. J. Hum. Genet.* **31**, 366–388 (1979).
16. L. L. Cavalli-Sforza, M. W. Feldman, The evolution of continuous variation. III. Joint transmission of genotype, phenotype and environment. *Genetics* **90**, 391–425 (1978).
17. D. C. Rao, N. E. Morton, J. M. Lalouel, R. Lew, Path analysis under generalized assortative mating: II. American IQ. *Genet. Res.* **39**, 187–198 (1982).
18. T. A. Manolio *et al.*, Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
19. M. W. Feldman, S. Ramachandran, Missing compared to what? Revisiting heritability, genes and culture. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, 20170064 (2018).
20. R. E. Nisbett *et al.*, Intelligence: New findings and theoretical developments. *Am. Psychol.* **67**, 130–159 (2012).
21. P. Wainschtein *et al.*, Recovery of trait heritability from whole genome sequence data. [BioRxiv:10.1101/588020](https://doi.org/10.1101/588020) (25 March 2019).
22. J. Yang *et al.*, Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
23. O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1193–1198 (2012).
24. A. A. E. Vinkhuyzen, S. Van Der Sluis, H. H. M. Maes, D. Posthuma, Reconsidering the heritability of intelligence in adulthood: Taking assortative mating and cultural transmission into account. *Behav. Genet.* **42**, 187–198 (2012).
25. F. C. Tropf *et al.*, Hidden heritability due to heterogeneity across seven populations. *Nat. Hum. Behav.* **1**, 757–765 (2017).
26. A. Kong *et al.*, The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018).
27. A. I. Young *et al.*, Relatedness disequilibrium regression estimates heritability without environmental bias. *Nat. Genet.* **50**, 1304–1310 (2018).
28. A. I. Young, Solving the missing heritability problem. *PLoS Genet.* **15**, e1008222 (2019).
29. A. I. Young, S. Benonisdottir, M. Przeworski, A. Kong, Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).
30. K. G. Ardlie, K. L. Lunetta, M. Seielstad, Testing for population subdivision and association in four case-control studies. *Am. J. Hum. Genet.* **71**, 304–311 (2002).
31. M. R. Robinson *et al.*, Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* **1**, 0016 (2017).
32. J. Pingault *et al.*, Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.* **19**, 566–580 (2018).
33. D. A. Briley, J. Livengood, J. Derringer, Behaviour genetic frameworks of causal reasoning for personality psychology. *Eur. J. Pers.* **32**, 202–220 (2018).
34. D. Conley, S. Zhang, The promise of genes for understanding cause and effect. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5626–5628 (2018).
35. T. T. Morris, N. M. Davies, G. Hemani, G. D. Smith, Population phenomena inflate genetic associations of complex social traits. *Sci. Adv.* **6**, eaay0328 (2020).
36. D. S. Falconer, T. F. C. Mackay, *Introduction to Quantitative Genetics* (Longman, Essex, UK, 1996).
37. M. Lynch, B. Walsh, *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, MA, 1998).
38. J. Pearl, Interpretation and identification of causal mediation. *Psychol. Methods* **19**, 459–481 (2014).
39. J. Pearl, *Causality: Models, Reasoning and Inference* (Cambridge University Press, ed. 2, 2009).
40. I. Shpitser, J. Pearl, Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.* **9**, 1941–1979 (2008).
41. J. Tian, J. Pearl, “A general identification condition for causal effects” in *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, R. Dechter, R. Sutton, program co-chairs (AAAI Press/MIT Press, Menlo Park, CA, 2002), pp. 567–573.
42. I. Shpitser, J. Pearl, “Identification of joint interventional distributions in recursive semi-Markovian causal models” in *Proceedings of the National Conference on Artificial Intelligence*, Y. Gil, R. J. Mooney, co-chairs (AAAI, Boston, MA, 2006), vol. 2, pp. 1219–1226.
43. C. Avin, I. Shpitser, J. Pearl, “Identifiability of path-specific effects” in *Proceedings of Nineteenth International Joint Conference on Artificial Intelligence*, L. P. Kaelbling, program chair (Morgan Kaufmann Publishers, San Francisco, CA, 2005), pp. 357–363.
44. I. Shpitser, T. J. VanderWeele, A complete graphical criterion for the adjustment formula in mediation analysis. *Int. J. Biostat.* **7**, 16 (2011).
45. H. Mostafavi *et al.*, Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9**, e48376 (2020).