Assortative Mating on Blood Type: Evidence from One Million Chinese Pregnancies[1]

Yao Hou[2], Ke Tang*[3], Jingyuan Wang*[4], Danxia Xie*[5], Hanzhe Zhang*[6]

October 16, 2022

Classification: Social Sciences/Social Sciences

Keywords: Assortative mating, blood type, mate choice

---

[1] All authors contributed equally and are ordered alphabetically. Correspondence should be addressed to: ketang@tsinghua.edu.cn, jywang@buaa.edu.cn, xiedanxia@tsinghua.edu.cn and hanzhe@msu.edu.

[2] houy20@mails.tsinghua.edu.cn; Institute of Economics, School of Social Sciences, Tsinghua University, Beijing, China. ORCID: 0000-0002-8818-4566

[3] ketang@tsinghua.edu.cn; Institute of Economics, School of Social Sciences, Tsinghua University and Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing, China.

[4] jywang@buaa.edu.cn; Beihang University, Beijing, China. ORCID: 0000-0003-0651-1592

[5] xiedanxia@tsinghua.edu.cn; Institute of Economics, School of Social Sciences, Tsinghua University, Beijing, China.

[6] hanzhe@msu.edu; Department of Economics, Michigan State University, East Lansing, Michigan, United States.

Abstract: Blood type is one of the most fundamental phenotypes in biological, medical, and psychological studies. Using a unique dataset of one million Chinese pregnancies, we find strong evidence from a group of statistical tests for assortative mating on blood type. After controlling for some anthropometric and socioeconomic confounders, assortative mating remains robust.

Significance Statement: In the human population, spousal pairs have been found to share phenotypes, which demonstrates the highly nonrandom nature of human mate choice. However, assortative mating on blood type—one of the most fundamental phenotypes in biological, medical, and psychological studies—has not been investigated. Using a unique dataset from China, we provide the first statistical analysis to test whether matching on blood type is nonrandom and find the first set of strong evidence for assortative mating on blood type. The findings are robust after we control for the effect of other possible mechanisms, such as population stratification, and show that the spousal concordance on blood type we observe is attributable to not only an individual's mate opportunity but also their mate choice.

Section 1. Introduction

Spousal pairs have been found to share a wide array of biological, socioeconomic, and psychological traits. One possible explanation is the highly nonrandom nature of human mate choice.[1-12] Assortative mating may bring about important genetic consequences by increasing trait variance in a population, intensifying trait divergence, and providing lifesaving benefits; for example, during organ transplants.[13-15] Motivated by significant multidisciplinary implications, studies reveal that many phenotypes, including body mass index (BMI), body height, and intelligence quotient (IQ),[10,16-21] are highly correlated within spousal pairs. However, blood type—one of the most fundamental phenotypes—was discovered more than a century ago but remains untapped in studies on assortative mating.[22] It is therefore of significance to investigate whether there is assortative mating on blood type in the population.

In this study, we aim to investigate whether there is assortative mating on blood type and investigate possible sources of spousal concordance. To this end, we utilize a dataset obtained from 2014-2015 Chinese free prepregnancy checkup, a project that targets low-income married couples in urban areas and married couples in rural areas who plan to get pregnant in the next 6 months. With a full sample of 931,964 pregnant couples, the dataset represents an unbiased sample of the target population. Two dimensions of selection left unaccounted for are nonmarital fertility and marital nonfertility, but both are very uncommon in China: For example, the proportion of children conceived by the 1970-1979 birth cohort out of marriage is 0.3%, according to 2016 China Family Panel Studies. According to 2017 National Fertility Survey of China, the fraction of nonmarital fertility is 7.9%. The survey also shows that only 1.5% of married women aged between 15 and 60 had not got pregnant.

First, we perform Pearson's chi-square test on blood-type associations within spousal pairs to evaluate the degree of assortative mating on blood type. Our dataset represents an unbiased sample of the population. Second, we use a series of alternative measures to check whether the obtained results are robust. Third, we perform meta-analysis using subsamples divided according to the regions in which couples received prepregnancy checkups to control for the effect of population stratification on the estimation. Next, we perform logistic regression analysis and linear regression analysis to isolate our estimates from population stratification or the relationship-maintenance effect. Specifically, we regress an individual's blood type on spousal blood type and incorporate the share of blood type in the local population, share of the individual's blood type in her ethnicity, and length of marriage in the regressions. Finally, we examine bivariate correlation between blood type and other phenotypes and perform mediation analysis to explore potential reasons for observed assortative mating on blood type.

Section 2. Results

We perform a series of statistical tests to explore assortative mating on blood type. We first use Pearson's chi-square test on a contingency table for spousal pairs' blood types to evaluate whether blood type influences human mate choice. Assortative mating typically refers to a mating choice pattern in which individuals with similar phenotypes mate with each other more frequently than the theoretical prediction under a random mating pattern. This definition implies that the degree of assortative mating on a certain phenotype can be measured by a chi-square test that compares the contingency table for pairs' blood types with the contingency

table generated by random matching. To perform this test, we use the full sample of Chinese prepregnancy checkup data and aggregate information on each spousal pair's blood types into a contingency table. The raw dataset consists of 1,137,010 couples who were followed and became pregnant within 6 months after the prepregnancy exam, from all 31 provincial administrative regions of Mainland China. For convenience for our later analysis, we remove observations with incomplete information related to the couple's blood types, living areas, birthplaces, ethnicity, and marital status. We obtain a sample of 931,964 couples with complete personal information to serve as our full sample. As the contingency table reports, diagonal elements have a higher frequency than expected by uniformly random mating, which shows that spousal pairs with the same blood type are more likely to marry each other (Table 1). Pearson's chi-square test on the full sample (chi-square statistic: 4020.942, p-value: 0.000, degree of freedom: 9, Cramer's V: 0.038) further validates nonrandom mating.

Table 1 Observed mating pairs and ratios over expected mating pairs

		Female			
		A	B	AB	O
Male	A	88,168 +5.2%	77,223 +0.1%	24,426 +0.1%	86,224 -4.9%
	B	79,506 +2.2%	75,959 +6.1%	22,842 +0.9%	77,904 -7.4%
	AB	24,277 -1.8%	24,033 +5.7%	8,492 +18.1%	24,564 -8.1%
	O	91,135 -5.8%	83,310 -6.4%	26,584 -5.5%	117,317 +12.2%

Note: Percentages of observed mating pairs above the expected mating pairs are underlined.

We subsequently adopt alternative measures to evaluate the degree of assortative mating to investigate whether the conclusion obtained from the chi-square test is robust. The chi-square test shows whether we can reject the null hypothesis of random mating, but it cannot tell us the specific pattern the mating process follows if the null hypothesis is rejected. To learn whether mating choice is assortative, we are particularly interested in the diagonal elements of the contingency table. Specifically, we want to know whether the numbers for matching in these four specific cells are significantly different from expected under uniformly random matching. To do this, we compute adjusted Pearson residuals that indicate whether the number of mating pairs in a specific cell is significantly different than expected and perform statistical tests on the statistics of diagonal elements at the $\alpha = 5\%$ level. It shows that the four pairs of blood types—i.e., the diagonal elements of the country-level contingency table, (A, A), (B, B), (AB, AB), and (O, O)—have adjusted Pearson residuals of 21.312, 22.420, 16.846, and 59.482, respectively, all of which are significantly higher than the two-sided critical value after Bonferroni correction, ± 2.498. The results suggest that individuals of all blood types tend to marry those who share the same blood type.

We also use the Altham index, which is widely used in testing the association of unordered rows and columns of an $r \times s$ contingency table, to alternatively measure the overall extent of assortative mating.[23-25] The Altham index uses the odds ratio of the likelihood of matching

within different blood type pairs to capture the distance between the row-column associations in the observed contingency table and those generated by random matching patterns (see Methods). Its value is equal to zero if mating choice is random and increases with the degree of nonrandom row-column association. The estimate of the Altham index with the full sample is 2.063, which demonstrates a significant correlation in spousal blood types.
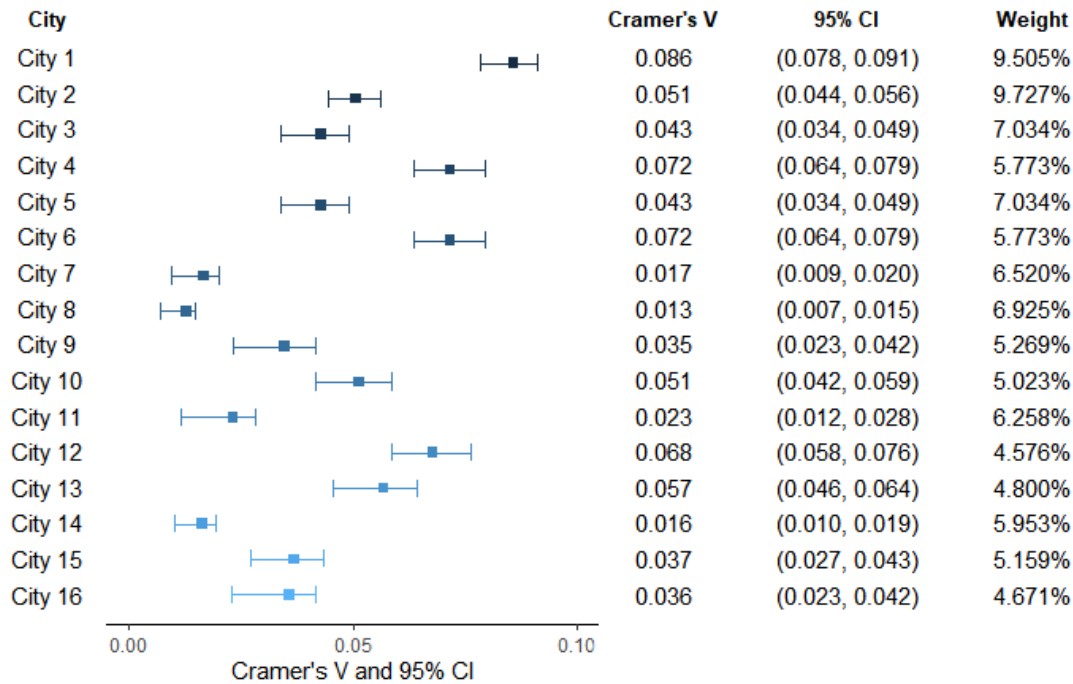
Because of possible geographical heterogeneity of blood type distribution, we then restrict our analysis to local subsamples (couples who were born in the same region and received a prepregnancy checkup in this region) and perform a meta-analysis on assortative mating to investigate whether the assortative mating pattern still exists after controlling for possible subpopulation structure, and whether this pattern is universal among different areas in China.[15] Population stratification is one of the key drivers of spousal concordance that is independent of individuals' mate choice. Individuals from subpopulations with residential proximity may naturally have more opportunities to mate. Moreover, they usually share similarities in their blood types because of their similar ethnic backgrounds. Without controlling for population stratification, we may overestimate the degree of assortative mating. Local subsample analysis is therefore proposed to relieve the estimation bias of assortative mating induced by population stratification. We identify locally matched observations by birthplace information and living area information provided by Chinese prepregnancy checkup data. Meta-analysis is carried out at city level (that is, we evaluate assortative mating within a city). We report statistical results for the 16 cities whose sample sizes (i.e., number of pregnant couples) are larger than 10,000 (Table 2, Figure 1).

Table 2 Statistical tests at city level (16 cities with samples>10,000)

Sample Size	Pearson's Chi-square statistic	Adjusted Pearson Residuals			
		A, A	B, B	AB, AB	O, O
34,015	751.419**	16.367**	8.858**	5.768**	25.060**
29,955	230.625**	1.826	-5.149**	-0.441	10.506**
22,124	121.338**	0.250	-4.493**	1.758	6.147**
18,662	287.433**	7.720**	7.138**	0.212	10.398**
18,178	15.148	1.167	1.309	-0.385	3.716**
17,045	8.244	1.142	0.822	-0.286	-0.402
16,845	61.116**	2.321	2.878**	1.575	7.171**
16,668	131.567**	5.810**	2.487	4.539**	8.346**
16,284	26.275**	3.134**	0.844	1.471	3.041**
15,550	213.853**	-9.092**	-2.533**	0.787	5.925**
15,530	150.266**	-3.102**	-1.402	6.452**	-0.147
15,062	12.042	1.481	0.610	1.686	1.400
14,956	61.087**	-2.171	-3.215**	2.567**	3.299**
14,349	54.662**	2.003	4.212**	2.714**	5.527**
12,922	41.119**	4.558**	0.826	2.505**	4.362**
11,409	255.578**	-4.481**	-5.478**	1.079	10.643**

Note: ** Significant at the 5% level. The 5%-level critical value for Pearson's chi-square test is 19.023 with a degree of freedom of 9. The 5%-level critical value for adjusted Pearson residuals with Bonferroni correction is 2.498 (see Methods).

Figure 1 Forest plot of meta-analysis

City		Cramer's V	95% CI	Weight
City 1		0.086	(0.078, 0.091)	9.505%
City 2		0.051	(0.044, 0.056)	9.727%
City 3		0.043	(0.034, 0.049)	7.034%
City 4		0.072	(0.064, 0.079)	5.773%
City 5		0.043	(0.034, 0.049)	7.034%
City 6		0.072	(0.064, 0.079)	5.773%
City 7		0.017	(0.009, 0.020)	6.520%
City 8		0.013	(0.007, 0.015)	6.925%
City 9		0.035	(0.023, 0.042)	5.269%
City 10		0.051	(0.042, 0.059)	5.023%
City 11		0.023	(0.012, 0.028)	6.258%
City 12		0.068	(0.058, 0.076)	4.576%
City 13		0.057	(0.046, 0.064)	4.800%
City 14		0.016	(0.010, 0.019)	5.953%
City 15		0.037	(0.027, 0.043)	5.159%
City 16		0.036	(0.023, 0.042)	4.671%

Cramer's V and 95% CI

Note: The standard error and 95% CI of Cramer's V are estimated through bootstrap method with 1,000 replications. The weight assigned to each subsample is equal to the inverse of the error variance of the subsample's Cramer's V over the sum of the error variance's inverses among subsamples. (see Methods)

In 13 of the 16 cities, Pearson's chi-square statistics indicate that mating choice is nonrandom at the 5% level of significance. Adjusted Pearson residuals further suggest that assortative mating on blood type is common in different areas, even after the subpopulation structure is controlled for. Blood-type pairs (O, O) typically have higher adjusted Pearson residuals than other blood type pairs, and thus show a higher degree of assortative mating. Also, the values of adjusted Pearson residuals we obtain from locally matched samples in each city are generally lower than those obtained from the national full sample. Similar pattern is observed in the forest plot, in which 12 of the 16 cities have an effect size, i.e., Cramer's V, larger than 0.03 and 7 among them have an effect size greater than 0.05 (Figure 1). The fixed effects estimate of Cramer's V over the 16 subsamples is 0.046.

Meta-analysis helps to eliminate the possible effects of subpopulation structure on assortative mating. However, another possible mechanism, relationship maintenance, may also confound estimation of assortative mating. The similarity of spousal pairs' blood types could be explained by the concordance of living with a partner with the same blood type. As some studies have indicated, spousal pairs with similar phenotypes, such as alcohol consumption, could be more likely to remain in a relationship.[26-28] If blood type is associated with some phenotypes, then couples with the same blood type might live in concord with each other because of the similarities they share and enroll in the prepregnancy checkup dataset, because their relationship has a lower probability of breaking up before they marry. Therefore, we perform regression analysis to isolate the effect of assortative mating on spousal concordance on blood type from that of relationship maintenance as well as that of subpopulation structure. The regressions will help us quantitatively evaluate how much variation in an individual's

blood type can be explained by assortative mating and the other two alternative mechanisms—population stratification and relationship maintenance. Specifically, we regress an individual's blood type on their partner's blood type and incorporate a group of control variables in the regression, which includes the share of the individual's blood type in the population in their birthplace, the share in their ethnicity, the length of the marriage, and its interaction term with the partner's blood type (see Methods for details). For comparison, we also run regressions without controlling for other factors.

The results are reported in Tables 3 and 4. In the odd-numbered columns, positive coefficients of indicator variables that show whether the individual and her partner have the same blood type offer evidence for assortative mating on blood type. After incorporating control variables, as reported in even-numbered columns, the magnitudes of indicator variables' coefficients decline considerably, and most control variables show statistical significance; this validates that estimates of assortative mating can be biased by confounding factors such as population stratification and relationship maintenance. It can be further seen from the decline in estimates of indicator variables' coefficients after incorporation of control variables that approximately 30% to 40% of spousal concordance on blood type observed in the odd-numbered columns in Table 3 and 4 is attributed to confounders, while the remaining portion of spousal similarities can be ascribed to assortative mating. After incorporation of a group of control variables, the coefficients of the partner's blood type are still statistically significant at the 1% level, which indicates highly nonrandom matching on blood type.

The coefficients of variables with statistical significance represent the extent to which the corresponding mechanism can explain the individual's blood type. By decomposing spousal concordance on blood type into several mechanisms, even-numbered columns in Tables 3 and 4 reflect the magnitude of different mechanisms' effects on spousal similarities observed in data. The results show that both subpopulation structure and assortative mating play key roles in explaining the observed pattern, while relationship maintenance fails to be a convincing explanation. One's blood type is mostly explained by the distribution of blood types in the population of her birthplace. A 1% rise in the share increases the odds ratio of having the corresponding blood type by approximately 4% to 5% for individuals with type A, B, or O blood and by approximately 12% for those with type-AB blood. In addition to local population structure, the individual's ethnicity explains a considerable fraction of her blood type. A 1% increase in the share of her blood type in her ethnicity increases the odds ratio of having the corresponding blood type by approximately 1-2% for those with type A, B, or O blood and by approximately 5% for those with type-AB blood. Her partner's blood type partner is also an effective predictor of her blood type, which provides evidence for assortative mating. If her partner has a given blood type, the odds ratio of having the same blood type will increase by approximately 8% for those with type-A blood, 7% for those with type-B blood, 15% for those with type-AB blood, and 20% for those with type-O blood (Tables 3 and 4). More assortative mating is observed within spousal pairs with type-AB blood or type-O blood. The almost equal estimates in Tables 3 and 4 suggest that the degrees of assortative mating are similar between females and males. We do not find strong evidence to support the argument that the similarity of spousal pairs' blood types is associated with length of marriage, which is used as a proxy for relationship maintenance: The coefficients of the interaction terms of length of marriage and the indicator variables for partner's blood type are insignificant or of small magnitude (lower

than 0.001).

The estimated increase in the odds ratio roughly indicates the degree of assortative mating on blood type. To estimate the increase in the probability of matching between individuals that can be explicitly attributed to blood-type assortative mating, we repeat the above regression analysis using a linear regression model and report the results in Tables 5 and 6. Like the results obtained by logistic regression analysis, individuals with the same blood type have a significantly higher probability of matching with each other than those with different blood types. The results are robust after controlling for a group of control variables. If one's partner is of a specific blood type, the probability of having the same blood type will increase by approximately 1.6% for those with type-A blood, 1.3% for those with type-B blood, 1.2% for those with type-AB blood, and 4.4% for those with type-O blood (Tables 5 and 6). Again, linear regression analysis provides clear evidence to support assortative mating on blood type.

Table 3 Logistic regression analysis of the full sample: Wife's mate choice

Variable	I_{wife}^{A}		I_{wife}^{B}		I_{wife}^{AB}		I_{wife}^{O}	
I_{Husband}^{A}	0.104***	0.076***						
	(0.005)	(0.006)						
I_{Husband}^{B}			0.115***	0.066***				
			(0.005)	(0.006)				
I_{Husband}^{AB}					0.203***	0.145***		
					(0.012)	(0.014)		
I_{Husband}^{O}							0.273***	0.197***
							(0.005)	(0.005)
$\text{BS}_{\text{Wife}}^{A}$	4.610***							
	(0.058)							
$\text{BS}_{\text{Wife}}^{B}$			4.821***					
			(0.049)					
$\text{BS}_{\text{Wife}}^{AB}$					12.054***			
					(0.206)			
$\text{BS}_{\text{Wife}}^{O}$							4.238***	
							(0.035)	
$\text{EGS}_{\text{Wife}}^{A}$	1.804***							
	(0.176)							
$\text{EGS}_{\text{Wife}}^{B}$			0.567*					
			(0.316)					
$\text{EGS}_{\text{Wife}}^{AB}$					4.821***			
					(0.603)			
$\text{EGS}_{\text{Wife}}^{O}$							0.573***	
							(0.096)	
Length of Marriage	0.000		0.000**		0.000		-0.000***	
	(0.000)		(0.000)		(0.000)		(0.000)	
I_{Husband}^{A} ×Length of Marriage	-0.000							
	(0.000)							
I_{Husband}^{B} ×Length of Marriage			-0.000					
			(0.000)					
I_{Husband}^{AB} ×Length of Marriage					0.000*			
					(0.000)			
I_{Husband}^{O} ×Length of Marriage							0.000**	
							(0.000)	
Observation	931,964	931,964	931,964	931,964	931,964	931,964	931,964	931,964

Note: Robust standard errors are in parentheses. * Significant at 10% level, ** 5% level, *** 1% level. $I_{\text{Husband,i}}^{g}$ and $I_{\text{Wife,i}}^{g}$ indicate whether husband and wife of couple i have type-g blood, respectively. $\text{BS}_{\text{Wife,i}}^{g}$ and $\text{BS}_{\text{Husband,i}}^{g}$ are the share of type-g blood individuals in the population of the birthplace of the wife or the husband in couple i. $\text{EGS}_{\text{Wife,i}}^{g}$ and $\text{EGS}_{\text{Husband,i}}^{g}$ represent the share of individuals with type-g blood in the population of the ethnicity of the wife or the husband in couple i. Length of Marriage$_i$ (Unit: day) indicates the time gap between when couple i marries and when they receive a prepregnancy checkup. (The same notations apply to the tables below.)

Table 4 Logistic regression analysis of the full sample: Husband's mate choice

Variable	$I^A_{Husband}$		$I^B_{Husband}$		$I^{AB}_{Husband}$		$I^O_{Husband}$	
I^A_{Wife}	0.104***	0.076***						
	(0.005)	(0.006)						
I^B_{Wife}			0.115***	0.066***				
			(0.005)	(0.006)				
I^{AB}_{Wife}					0.203***	0.146***		
					(0.012)	(0.014)		
I^O_{Wife}							0.273***	0.195***
							(0.005)	(0.005)
$BS^A_{Husband}$		4.667***						
		(0.055)						
$BS^B_{Husband}$				4.898***				
				(0.047)				
$BS^{AB}_{Husband}$						11.598***		
						(0.172)		
$BS^O_{Husband}$								4.210***
								(0.035)
$EGS^A_{Husband}$		1.441***						
		(0.183)						
$EGS^B_{Husband}$				0.772***				
				(0.269)				
$EGS^{AB}_{Husband}$						5.005***		
						(0.623)		
$EGS^O_{Husband}$								0.482***
								(0.093)
Length of Marriage		0.000		-0.000		-0.000*		0.000
		(0.000)		(0.000)		(0.000)		(0.000)
I^A_{Wife} ×Length of Marriage		-0.000						
		(0.000)						
I^B_{Wife} ×Length of Marriage				-0.000				
				(0.000)				
I^{AB}_{Wife} ×Length of Marriage						0.000*		
						(0.000)		
I^O_{Wife} ×Length of Marriage								0.000**
								(0.000)
Observation	931,964	931,964	931,964	931,964	931,964	931,964	931,964	931,964

Note: Robust standard errors are in parentheses. * Significant at 10% level, ** 5% level, *** 1% level. $I^g_{Husband,i}$ and $I^g_{Wife,i}$ indicate whether husband and wife of couple i have type-g blood, respectively. $BS^g_{Wife,i}$ and $BS^g_{Husband,i}$ are the share of type-g blood individuals in the population of the birthplace of the wife or the husband in couple i. $EGS^g_{Wife,i}$ and $EGS^g_{Husband,i}$ represent the share of individuals with type-g blood in the population of the ethnicity of the wife or the husband in couple i. Length of Marriage$_i$ (Unit: day) indicates the time gap between when couple i marries and when they receive a prepregnancy checkup.

Table 5 Linear regression analysis of the full sample: Wife's mate choice

| Variable	I^A_{wife}		I^B_{wife}		I^{AB}_{wife}		I^O_{wife}	
I^A_{Husband}	0.022***	0.016***						
	(0.001)	(0.001)						
I^B_{Husband}			0.023***	0.013***				
			(0.001)	(0.001)				
I^{AB}_{Husband}					0.018***	0.012***		
					(0.001)	(0.001)		
I^O_{Husband}							0.061***	0.044***
							(0.001)	(0.001)
$\text{BS}^A_{\text{Wife}}$		0.961***						
		(0.012)						
$\text{BS}^B_{\text{Wife}}$				0.984***				
				(0.010)				
$\text{BS}^{AB}_{\text{Wife}}$						0.963***		
						(0.015)		
$\text{BS}^O_{\text{Wife}}$								0.941***
								(0.008)
$\text{EGS}^A_{\text{Wife}}$		0.286***						
		(0.031)						
$\text{EGS}^B_{\text{Wife}}$				0.114*				
				(0.064)				
$\text{EGS}^{AB}_{\text{Wife}}$						0.236***		
						(0.037)		
$\text{EGS}^O_{\text{Wife}}$								0.167***
								(0.023)
Length of Marriage		0.000		0.000**		0.000		-0.000***
		(0.000)		(0.000)		(0.000)		(0.000)
$I^A_{\text{Husband}} \times$ Length of Marriage		-0.000						
		(0.000)						
$I^B_{\text{Husband}} \times$ Length of Marriage				-0.000				
				(0.000)				
$I^{AB}_{\text{Husband}} \times$ Length of Marriage						0.000**		
						(0.000)		
$I^O_{\text{Husband}} \times$ Length of Marriage								0.000
								(0.000)
Observation	931,964	931,964	931,964	931,964	931,964	931,964	931,964	931,964

Note: Robust standard errors are in parentheses. * Significant at 10% level, ** 5% level, *** 1% level. $I^g_{\text{Husband},i}$ and $I^g_{\text{Wife},i}$ indicate whether husband and wife of couple i have type-g blood, respectively. $\text{BS}^g_{\text{Wife},i}$ and $\text{BS}^g_{\text{Husband},i}$ are the share of type-g blood individuals in the population of the birthplace of the wife or the husband in couple i. $\text{EGS}^g_{\text{Wife},i}$ and $\text{EGS}^g_{\text{Husband},i}$ represent the share of individuals with type-g blood in the population of the ethnicity of the wife or the husband in couple i. Length of Marriage$_i$ (Unit: day) indicates the time gap between when couple i marries and when they receive a prepregnancy checkup.

Table 6 Linear regression analysis of the full sample: Husband's mate choice

Variable	$I^A_{Husband}$		$I^B_{Husband}$		$I^{AB}_{Husband}$		$I^O_{Husband}$	
I^A_{Wife}	0.022***	0.016***						
	(0.001)	(0.001)						
I^B_{Wife}			0.023***	0.013***				
			(0.001)	(0.001)				
I^{AB}_{Wife}					0.017***	0.012***		
					(0.001)	(0.001)		
I^O_{Wife}							0.062***	0.044***
							(0.001)	(0.001)
$BS^A_{Husband}$		0.970***						
		(0.011)						
$BS^B_{Husband}$				0.986***				
				(0.009)				
$BS^{AB}_{Husband}$						0.977***		
						(0.015)		
$BS^O_{Husband}$								0.946***
								(0.008)
$EGS^A_{Husband}$		0.217***						
		(0.032)						
$EGS^B_{Husband}$				0.144***				
				(0.054)				
$EGS^{AB}_{Husband}$						0.181***		
						(0.037)		
$EGS^O_{Husband}$								0.148***
								(0.022)
Length of Marriage		0.000		-0.000		- 0.000***		0.000
		(0.000)		(0.000)		(0.000)		(0.000)
I^A_{Wife} ×Length of Marriage		-0.000						
		(0.000)						
I^B_{Wife} ×Length of Marriage				-0.000				
				(0.000)				
I^{AB}_{Wife} ×Length of Marriage						0.000*		
						(0.000)		
I^O_{Wife} ×Length of Marriage								0.000**
								(0.000)
Observation	931,964	931,964	931,964	931,964	931,964	931,964	931,964	931,964

Note: Robust standard errors are in parentheses. * Significant at 10% level, ** 5% level, *** 1% level. $I^g_{Husband,i}$ and $I^g_{Wife,i}$ indicate whether husband and wife of couple i have type-g blood, respectively. $BS^g_{Wife,i}$ and $BS^g_{Husband,i}$ are the share of type-g blood individuals in the population of the birthplace of the wife or the husband in couple i. $EGS^g_{Wife,i}$ and $EGS^g_{Husband,i}$ represent the share of individuals with type-g blood in the population of the ethnicity of the wife or the husband in couple i. Length of Marriage$_i$ (Unit: day) indicates the time gap between when couple i marries and when they receive a prepregnancy checkup.

Section 3. Possible Reasons for Assortative Mating on Blood Type

Having shown robust evidence for assortative mating on blood type, we investigate potential reasons. One possible explanation is that blood type may act as a proxy for other phenotypes. As previously stated, many studies have validated assortative mating on a group of phenotypes, such as BMI, weight, height, and IQ.[10,16-21] That is, individuals tend to choose a partner who shares similarities along these dimensions when making mate choices. If blood type is associated with these phenotypes, spousal concordance on blood type will be observed because of assortative mating. Using personal information provided by the dataset, we examine bivariate correlation between blood type and other phenotypes (Figure 2). There appear to be some associations between blood type and the phenotypes we examine: education, job type, height, weight, pressure, and drinking habits. However, most associations have a relatively small correlation coefficient lying between -0.03 and 0.03.

To further explore to what extent assortative mating on blood type can be explained by its correlation with other phenotypes, we perform mediation analysis. Specifically, we first regress the individual's blood type on her partner's using a logistic regression model, then incorporate a mediator—i.e., one of the partner's phenotypes that might be associated with his blood type—to see whether and to what degree the effect of the partner's blood type on the individual's blood type is weakened after the mediator is included in the regression. We report the results of mediation analysis in Table 7 and 8. As can be seen, the coefficients of the partner's blood type decline after we include different mediators in the regression models, which shows that the associations between blood type and other phenotypes can explain assortative mating on blood type to a certain degree. We see from Columns 2-9 in Table 7 and Column 1 in Table 8 that the proportion of the coefficients of the partner's blood type absorbed by mediators varies with blood type. For individuals with type-B blood, when all mediators are included, the coefficients of the partner's blood type are reduced by around 6% to 7%; for those with type-A blood, the incorporation of mediators has little effect on estimation results for coefficients of the partner's blood type, as shown in Column 1 in Table 8. As for those with type-AB blood or type-O blood, the scale of mediator absorption is about 3% to 4%. However, a large fraction of assortativity remains unexplained. When we further include a group of control variables to isolate our measure of assortative mating from confounding factors—such as population stratification, province-level fixed effects, or even the individual's phenotypes—in the regression, as indicated by the statistical significance of the coefficients of the partner's blood type in Columns 2-4 in Table 8, we still find strong evidence for assortative mating on blood type. These findings suggest that there could be other potential mechanisms for this pattern we observe in the data. Further investigation of this is left for future research.

Figure 2 Bivariate correlation between blood type and other phenotypes
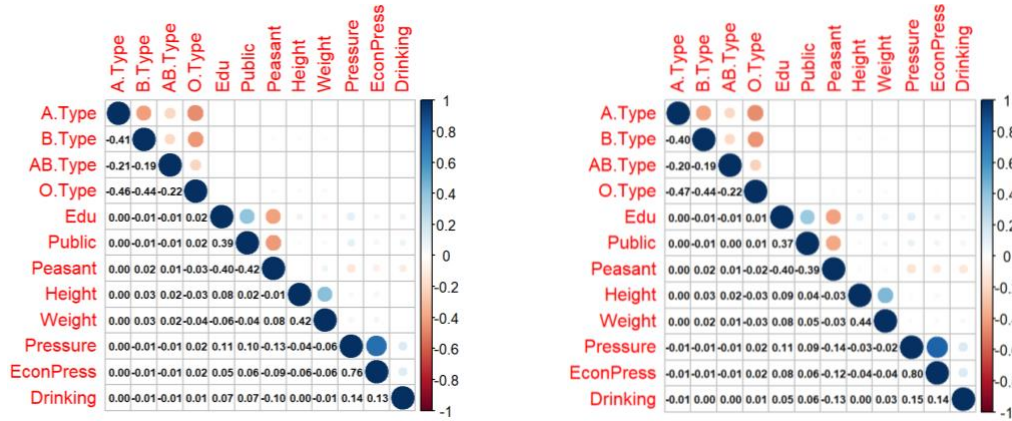
Wife's correlation matrix   Husband's correlation matrix

Note. The variables "x.Type" indicate whether the individual has type-x blood. Edu is a categorical variable that takes the value 1 if the individual receives no education, 2 if primary-school education, 3 if middle-school education, 4 if high-school education, 5 if a bachelor's degree, and 6 if postgraduate qualification. Public is a binary variable that indicates whether the individual works as a public employee. Peasant is a binary variable that represents if the individual is a peasant. Height (Unit: cm) and Weight (Unit: kg) record the individual's height and weight when receiving the prepregnancy examination. Pressure is a categorical variable that documents the degree of pressure felt by the individual and takes the value 0 if they feel no pressure, 4 if the pressure is heavy, and a discrete value of 1, 2, or 3 if the degree of pressure is moderate. EconPress is a categorical variable that measures the individual's economic pressure and is defined in a way that is similar to Pressure. Drinking is a categorical variable that rates how often the individual drinks and takes the value 0 if they never drink, 1 if they occasionally drink, and 2 if they often drink.

Table 7 Mediation analysis with single mediators

Mediator	No mediators	Edu	Public	Peasant	Height	Weight	Pressure	EconPress	Drinking
$I_{\text{Wife}}^{\text{A}}$	0.104***	0.104***	0.104***	0.104***	0.104***	0.104***	0.104***	0.104***	0.104***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$I_{\text{Wife}}^{\text{B}}$	0.115***	0.114***	0.114***	0.113***	0.112***	0.113***	0.114***	0.114***	0.115***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$I_{\text{Wife}}^{\text{AB}}$	0.203***	0.203***	0.203***	0.202***	0.200***	0.201***	0.202***	0.203***	0.203***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
$I_{\text{Wife}}^{\text{O}}$	0.273***	0.272***	0.273***	0.270***	0.269***	0.271***	0.272***	0.272***	0.273***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$I_{\text{Husband}}^{\text{A}}$	0.104***	0.104***	0.104***	0.104***	0.104***	0.104***	0.104***	0.104***	0.104***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$I_{\text{Husband}}^{\text{B}}$	0.115***	0.114***	0.114***	0.112***	0.111***	0.111***	0.114***	0.114***	0.115***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$I_{\text{Husband}}^{\text{AB}}$	0.203***	0.203***	0.203***	0.202***	0.200***	0.200***	0.202***	0.203***	0.203***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
$I_{\text{Husband}}^{\text{O}}$	0.273***	0.273***	0.273***	0.270***	0.268***	0.268***	0.272***	0.272***	0.273***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Controls	No	No	No	No	No	No	No	No	No
Regional fixed effect	No	No	No	No	No	No	No	No	No
Individual Phenotypes	No	No	No	No	No	No	No	No	No

Note: This table reports estimated coefficients of the partner's blood type in logistic regressions under different model settings. Column 1 reports estimation results in baseline logistic models that regress the individual's blood type on the partner's blood type: $ln[p_{\text{Wife,i}}^{\text{g}}/(1 - p_{\text{Wife,i}}^{\text{g}})] = \beta_0 + \beta_1 I_{\text{Husband,i}}^{\text{g}} + \epsilon_i, \text{g} \in \{A, B, AB, O\}$ and $ln[p_{\text{Husband,i}}^{\text{g}}/(1 - p_{\text{Husband,i}}^{\text{g}})] = \beta_0 + \beta_1 I_{\text{Wife,i}}^{\text{g}} + \epsilon_i, \text{g} \in \{A, B, AB, O\}$, where $p_{\text{Wife,i}}^{\text{g}}$ and $p_{\text{Husband,i}}^{\text{g}}$ indicate the probability that the wife or the husband of couple i has type-g blood. $I_{\text{Husband,i}}^{\text{g}}$ and $I_{\text{Wife,i}}^{\text{g}}$ refer to indicator variables that show whether the wife or the husband of couple i has type-g blood. Columns 2-9 report results of logistic models that incorporate a mediator: $ln[p_{\text{Wife,i}}^{\text{g}}/(1 - p_{\text{Wife,i}}^{\text{g}})] = \beta_0 + \beta_1 I_{\text{Husband,i}}^{\text{g}} + \beta_2 \text{Mediator}_{\text{Husband,i}} + \epsilon_i, \text{g} \in \{A, B, AB, O\}$ and $ln[p_{\text{Husband,i}}^{\text{g}}/(1 - p_{\text{Husband,i}}^{\text{g}})] = \beta_0 + \beta_1 I_{\text{Wife,i}}^{\text{g}} + \beta_2 \text{Mediator}_{\text{Wife,i}} + \epsilon_i, \text{g} \in \{A, B, AB, O\}$, where $\text{Mediator}_{\text{Husband,i}}$ and $\text{Mediator}_{\text{Wife,i}}$ measure the phenotype of the wife or the husband of couple i.

Table 8 Mediation analysis with all mediators

Mediator	All mediators	All mediators	All mediators	All mediators
I_{Wife}^{A}	0.104***	0.076***	0.076***	0.076***
	(0.005)	(0.006)	(0.006)	(0.006)
I_{Wife}^{B}	0.108***	0.066***	0.066***	0.066***
	(0.005)	(0.006)	(0.006)	(0.006)
I_{Wife}^{AB}	0.197***	0.145***	0.144***	0.144***
	(0.012)	(0.014)	(0.014)	(0.014)
I_{Wife}^{O}	0.264***	0.197***	0.198***	0.198***
	(0.005)	(0.005)	(0.005)	(0.005)
$I_{Husband}^{A}$	0.104***	0.076***	0.076***	0.076***
	(0.005)	(0.006)	(0.006)	(0.006)
$I_{Husband}^{B}$	0.107***	0.066***	0.066***	0.066***
	(0.005)	(0.006)	(0.006)	(0.006)
$I_{Husband}^{AB}$	0.196***	0.146***	0.144***	0.144***
	(0.012)	(0.014)	(0.014)	(0.014)
$I_{Husband}^{O}$	0.263***	0.195***	0.196***	0.196***
	(0.005)	(0.005)	(0.005)	(0.005)
Controls	No	Yes	Yes	Yes
Regional fixed effect	No	No	Yes	Yes
Individual Phenotypes	No	No	No	Yes

Note: This table reports estimated coefficients of the partner's blood type in logistic regressions under different model settings. Column 1 reports results of the logistic model that incorporates all mediators—i.e., the partner's phenotypes examined in Columns 2-9 of Table 7. Column 2 reports results of the logistic model that incorporates all mediators examined in Columns 2-9 of Table 7 and additional control variables, including the share of the individual's blood type in the population living in her birthplace, the share in their ethnicity, length of marriage, and its interaction term with the partner's blood type. Column 3 shows results of logistic models with incorporation of all mediators examined in Column 2-9 of Table 7, control variables specified in the model used in Column 2 as well as provincial-level fixed effects. Finally, Column 4 adopts a logistic model that adds the individual's phenotypes that are examined in Column 2-9 of Table 7, into the model used in Column 3 (see Methods for details).

Section 4. Conclusion

In summary, we provide evidence of assortative mating on blood type. The degree of assortative mating varies among individuals with different blood types and across locations. Our findings are robust after we control for other possible mechanisms, such as environmental confounding and relationship maintenance. We further examine potential mechanisms for these observations.

Our study makes two contributions. First, our empirical results shed light on nonrandom matching on blood type—one of the most well-known human phenotypes—which, to the best of our knowledge, has not been fully investigated. Assortative mating on blood type will have important genetic consequences by influencing the direction of the evolution of blood-type distribution in the population. On the one hand, it will intensify trait divergence compared with random mating, and thus heighten the population's response to directional natural selection.[29,30] On the other hand, it reduces the heterozygosity of the population's blood type and may promote inbreeding depression.[31] These evolutionary effects render it important to investigate this issue.

Second, we improve the causal inference of assortative mating by using a group of approaches. To mitigate the estimation bias caused by population stratification, we restrict our analysis to locally matched subsamples to perform meta-analysis. We further address this concern by running regressions with control variables. Other possible mechanisms are also controlled for in our analysis. Our robust results show causal evidence for assortative mating on blood type.

We acknowledge the limitations of our study. First, couples who are lost to follow up or fail to become pregnant after the prepregnancy examination are missing from our dataset. Selection into the sample introduced by the two kinds of prescreening could be (but does not appear to be) a potential source of bias. Second, although overall evidence for assortative mating was found and some potential mechanisms behind the pattern were examined, a large fraction of the assortative mating we observe from the data remains unexplained. It would be of interest to understand the underlying mechanisms in future research. Third, we should be cautious when extrapolating findings in the Chinese population to other populations. Further evidence is needed to investigate whether the nonrandom matching pattern of blood types is also robust in other populations.

Finally, we sought to avoid our estimates of the degree of assortative mating from being confounded by other factors,[6,21,28,32-42] such as population stratification and relationship maintenance after spousal pairing. We acknowledge that it is difficult to infer a causal relationship between blood-type similarity and mate choice. As previously noted, when a latent subpopulation structure underlies the observed sample, the positive association of blood type within spousal pairs can be attributed to systematic differences among subpopulations that arise from the ancestral differences they inherit. Under this circumstance, assortative mating will thus be overestimated if homogeneous sampling is assumed. An alternative explanation for the similarity of blood types within spousal pairs is that blood-type concordance helps spousal pairs maintain their relationship. Partners with the same blood type might share similar phenotypes, and thus find it easier to get along and maintain their relationship, which increases their probability of being enrolled in the study and leads to biased estimation. All three possible mechanisms may result in spousal concordance on blood type (Figure 3). We strove to control for confounding factors by incorporating a group of control variables to indicate the blood-type

distribution of subpopulations and the length of couples' relationship when estimating the extent of assortative mating on blood type using regression models. However, there might still be concern regarding whether confounding factors have been effectively ruled out and a causal relationship between blood-type concordance and mate choice is clearly identified. To further study assortative mating on blood type in future research, we will need to carefully distinguish this mechanism from the other two possible explanations.

Figure 3 Possible mechanisms of spousal concordance on blood type
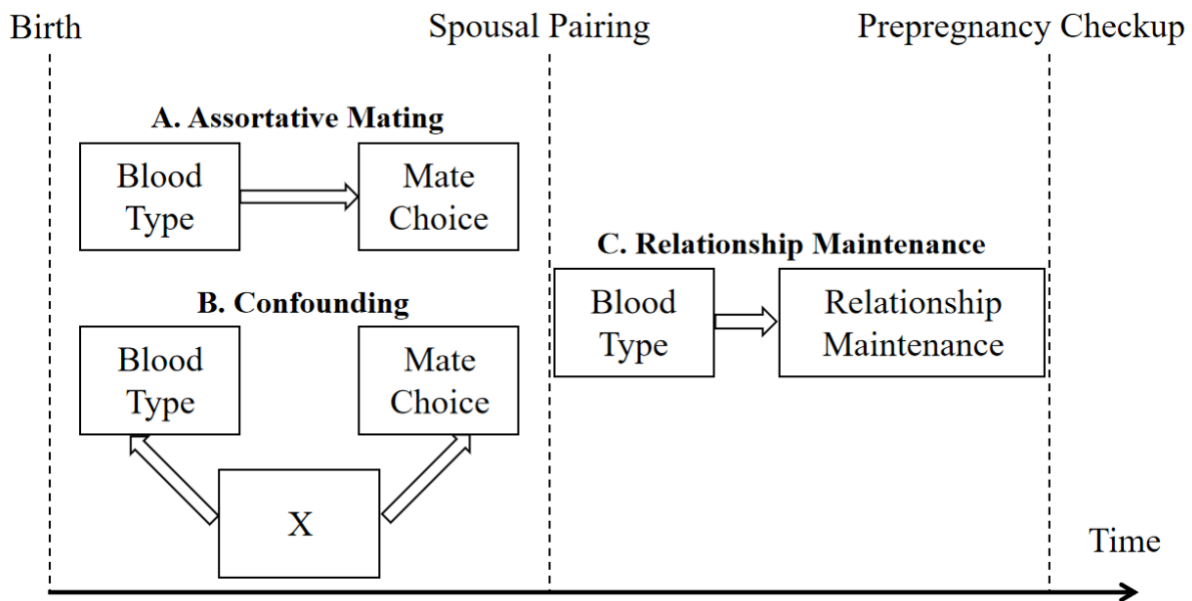
Note: Mechanism A (Assortative mating) refers to the process whereby partners choose each other based on their similarity in blood type. Mechanism B (Confounding) indicates confounding factors that may influence both blood type and mate choice; population stratification is the most common confounding factor. Mechanism C (Relationship Maintenance) reflects the possibility that spousal pairs with the same blood type are more likely to maintain a long-term relationship and be enrolled in the observed sample.

Section 5. Methods
Section 5A. Data
Chinese prepregnancy checkup data
In this study, we used 2014-2015 Chinese prepregnancy checkup data to perform statistical analysis. Our dataset is from the National Free Preconception Health Examination Project (NFPHEP), which was launched by the National Health Commission (NHC) and the Ministry of Finance of the People's Republic of China and offers free prepregnancy examinations for low-income married couples in urban areas and married couples in rural areas who plan to get pregnant in the next 6 months. The nationwide project covers married couples with pregnancy plans in 2,907 counties from all 31 provincial administrative regions of Mainland China. Aiming to reduce birth defects and improve the health of newborns, China's NHC strives to expand coverage of the survey with the aid of local communities and family planning service agencies. According to China's NHC, the survey covered over 95% of the targeted population between 2014 and 2015.[43,44] Couples who had made plans for pregnancy were enrolled by local community staff and given prepregnancy examinations. Blood samples are collected during the

examination and sent to local laboratories for blood-type testing. In addition to the results of physical examinations, the NFPHEP also collects all participants' basic personal information— e.g., sex, address of residence, ethnicity, birthplace, and marriage—via a standardized questionnaire as well as participants' identification cards.[45-47] Two follow-up telephone interviews are conducted after the examination by trained nurses. The first is performed within 3 months after the examination to check pregnancy status and the second is carried out within 1 year after the first follow-up interview to track the pregnancy outcome. After excluding couples who are lost to follow up or fail to become pregnant after the prepregnancy examination, 1,137,010 couples are included in the dataset. Informed consent is signed by all project participants before enrollment. This study was approved by NHC.

Subsample with complete personal information
We remove observations with incomplete information related to the couple's blood types, living areas, birthplaces, ethnicity, marriage, education, job type, body height, weight, pressure, and drinking habits, which we use in our statistical analysis, and obtain a sample of 931,964 couples for the full sample of our study. The raw dataset of Chinese prepregnancy checkup data was filtered in Stata.

Section 5B. Statistical analysis
Contingency table
We broke down the numbers of matches between individuals with different blood types by sex to produce a 4×4 contingency table, with the number in grid (i, j) $N_{i,j}$ representing the frequency of matching between males with type-i blood and females with type-j blood. The ratios of observed frequencies over expected frequencies are also reported, with gird (i, j) computed as $N_{i,j}/E_{i,j} = N_{i,j}/(R_i \times C_j/N)$, where R_i and C_j denote row and column marginal totals and N represents the grand total. The contingency table was produced using R.

Pearson's chi-square test
We performed a chi-square test on a contingency table of blood types obtained from the full sample to test the association between an individual's blood type and her partner's blood type, with a null hypothesis that matching behavior is independent of spousal pairs' blood types. The 5%-level critical value for Pearson's chi-square test is 19.023 with a degree of freedom of 9. Cramer's V was also reported to measure the effect size for the chi-square test. Pearson's chi-square test was performed using R.

Adjusted Pearson residuals
The adjusted Pearson residual of grid (i, j) is computed as

$$\tilde{r}_{i,j} = \frac{N_{i,j} - E_{i,j}}{\sqrt{E_{i,j}(1 - R_i/N)(1 - C_j/N)}} \tag{1}$$

It follows a standard normal distribution $N(0,1)$ in the null hypothesis of random matching, which enables us to perform statistical tests on nonrandom mating. The 5%-level critical value

19

for adjusted Pearson residuals is ±2.498. It is corrected by Bonferroni correction. To test the 4 diagonal elements in the contingency table, the new alpha level should be set as $\alpha_{Bon} = \frac{5\%}{4} = 1.25\%$ and the corresponding critical value $N(0,1)_{1-\alpha_{Bon}/2} = 2.498$. Computation of adjusted Pearson residuals and related statistical tests were performed using R.

The Altham index

We use the Altham index as an alternative measure of assortative mating, which can be written as

$$[\sum_{i \in G} \sum_{j \in G} \sum_{k \in G} \sum_{l \in G} |log(\frac{P(I^i_{Wife} = 1, I^j_{Husband} = 1)P(I^k_{Wife} = 1, I^l_{Husband} = 1)}{P(I^i_{Wife} = 1, I^l_{Husband} = 1)P(I^k_{Wife} = 1, I^j_{Husband} = 1)})|^2]^{1/2},$$
$$G = \{A, B, AB, O\} \tag{2}$$

where $I^g_{Husband}$ and I^g_{Wife} refer to indicator variables that show whether the wife or the husband has type-g blood, and $P(I^g_{Wife} = 1, I^h_{Husband} = 1)$ represents the probability of matching between females with type-g blood and males with type-h blood, which is estimated by the share of this type of matching in the full sample. The higher the index, the more nonrandom the mate choice; it is equal to zero when matching is random. The Altham index was computed using R.

Meta-analysis

As a sensitivity analysis, we restricted samples for Pearson's chi-square test to locally matched couples to prevent our test on assortative mating from being confounded by population stratification. Specifically, we cover couples who were born in the same birthplace and received prepregnancy checkups in this area and stratify them by their birthplace. Meta-analysis is performed at the city level, which allows more granular segmentation for 16 locally matched subsamples. The meta-analysis results of chi-square test and effect size (Cramer's V) estimation among subsamples are reported through a summary table and a forest plot. The standard error and 95% CI of Cramer's V are estimated through boostrap method with 1,000 replications. A fixed-effects meta-analysis model is further utilized to estimate the average true effect size of assortative mating on blood type over the 16 cities θ_{pop}. The model estimates it by computing the weighted average of true city-specific effect sizes θ_j $(j = 1,2,\ldots,16)$ that have been estimated y the proceeding meta-analysis,

$$\theta_{pop} = Ave(\theta_j) = \sum_{j=1}^{16} W_j \theta_j$$

in which the weight assigned to each subsample W_j is equal to the inverse of the error variance of the subsample's Cramer's V over the sum of the error variance's inverses among subsamples. The full sample was filtered in Stata to obtain locally matched subsamples. The bootstrap estimation of the standard error and 95% CI of Cramer's V among subsamples was also

20

processed in Stata. The rest parts of meta-analysis of subsamples were performed using R.

Regression Analysis

We investigated the degree of assortative mating on blood type with logistic regression models. The models are specified as equations (3) to (6), where $p_{\text{Wife,i}}^{\text{g}}$ and $p_{\text{Husband,i}}^{\text{g}}$ indicate the probability that the wife or the husband of couple i has type-g blood. $I_{\text{Husband,i}}^{\text{g}}$ and $I_{\text{Wife,i}}^{\text{g}}$ refer to indicator variables that show whether the wife or the husband of couple i has type-g blood. $\text{BS}_{\text{Wife,i}}^{\text{g}}$ and $\text{BS}_{\text{Husband,i}}^{\text{g}}$ suggest the share of individuals with type-g blood in the population of the birthplace of the wife or the husband in couple i, which is estimated by the population in this city covered in our full sample. $\text{EGS}_{\text{Wife,i}}^{\text{g}}$ and $\text{EGS}_{\text{Husband,i}}^{\text{g}}$ represent the share of individuals with type-g blood in the population of the ethnicity of the wife or the husband in couple i, which is estimated by the population in this ethnicity covered in our full sample. Finally, $\text{Length of Marriage}_{\text{i}}$ indicates the time duration between when couple i get married and when they receive a prepregnancy checkup. $I_{\text{Wife,i}}^{\text{g}} \times \text{Length of Marriage}_{\text{i}}$ and $I_{\text{Husband,i}}^{\text{g}} \times \text{Length of Marriage}_{\text{i}}$ are interaction terms of the indicator variables and the length of marriage of this couple. We also repeated the above steps using linear regression model, to explicitly estimate the increase in probability of matching induced by assortative mating on blood type.

$$ln\left(\frac{p_{\text{Wife,i}}^{\text{g}}}{1-p_{\text{Wife,i}}^{\text{g}}}\right) = \beta_0 + \beta_1 I_{\text{Husband,i}}^{\text{g}}, \qquad g \in \{A, B, AB, O\} \tag{3}$$

$$ln\left(\frac{p_{\text{Wife,i}}^{\text{g}}}{1-p_{\text{Wife,i}}^{\text{g}}}\right) = \beta_0 + \beta_1 I_{\text{Husband,i}}^{\text{g}} + \beta_2 \text{BS}_{\text{Wife,i}}^{\text{g}} + \beta_3 \text{EGS}_{\text{Wife,i}}^{\text{g}} + \beta_4 \text{Length of Marriage}_{\text{i}} +$$

$$\beta_5 I_{\text{Husband,i}}^{\text{g}} \times \text{Length of Marriage}_{\text{i}}, \qquad g \in \{A, B, AB, O\} \tag{4}$$

$$ln\left(\frac{p_{\text{Husband,i}}^{\text{g}}}{1-p_{\text{Husband,i}}^{\text{g}}}\right) = \beta_0 + \beta_1 I_{\text{Wife,i}}^{\text{g}}, \qquad g \in \{A, B, AB, O\} \tag{5}$$

$$ln\left(\frac{p_{\text{Husband,i}}^{\text{g}}}{1-p_{\text{Husband,i}}^{\text{g}}}\right)$$

$$= \beta_0 + \beta_1 I_{\text{Wife,i}}^{\text{g}} + \beta_2 \text{BS}_{\text{Husband,i}}^{\text{g}} + \beta_3 \text{EGS}_{\text{Husband,i}}^{\text{g}} + \beta_4 \text{Length of Marriage}_{\text{i}}$$

$$+$$

$$\beta_5 I_{\text{Wife,i}}^{\text{g}} \times \text{Length of Marriage}_{\text{i}}, \qquad g \in \{A, B, AB, O\} \tag{6}$$

Besides the examination of the degree of assortative mating via logistic regression models, we also perform mediation analysis to explore potential mechanisms behind the observed pattern. Specifically, we include different mediators the regression models in equations (3) and (5) and

estimate the decline of the coefficients of the partner's blood type that can be attributed to the incorporation of mediators, which is shown as equations (7) and (8).

$$ln\left(\frac{p_{Wife,i}^g}{1-p_{Wife,i}^g}\right) = \beta_0 + \beta_1 I_{Husband,i}^g + \beta_2 Mediator_{Husband,i}, \qquad g \in \{A, B, AB, O\} \qquad (7)$$

$$ln\left(\frac{p_{Husband,i}^g}{1-p_{Husband,i}^g}\right) = \beta_0 + \beta_1 I_{Wife,i}^g + \beta_2 Mediator_{Wife,i}, \qquad g \in \{A, B, AB, O\} \qquad (8)$$

where $Mediator_{Wife,i}$ and $Mediator_{Husband,i}$ measure the phenotype of the wife and the husband of couple i, respectively. The mediators we employ are education, whether the individual is a public employee, whether they work as a peasant, their height and weight, the pressure they feel, the economic pressure they feel, and their drinking habit. For all models, we adopt robust standard errors in the model estimation to ensure that our estimates are robust when the model specification is incorrect. Regression analysis was performed using Stata.

Section 5C. Data, Materials, and Software Availability
We use 2014-2015 Chinese prepregnancy checkup data available from the Institute of Science and Technology of the National Health Commission of the People's Republic of China. Data and code have been deposited in https://cloud.tsinghua.edu.cn/d/e86e227d8e66475ba790/.[48]

Table 9 List of variables used in analysis

Description	Value
Blood type	A, B, AB, O
Province the respondent currently lives in	province index
City the respondent currently lives in	city index
Province the respondent was born in	province index
City the respondent was born in	city index
Ethnic group	ethnic index
Duration between marriage registration and prepregnancy exam	Length (days)
Respondent's education level	1 to 6
Whether the respondent works as a public employee	binary
Whether the respondent is a peasant	binary
Respondent's height	in cm
Respondent's weight	in kg
The degree of pressure felt	0 to 4
The degree of economic pressure felt	0 to 4
How often the respondent drinks	0 to 2

Competing interests. The authors declare no competing interests.

References

1. Vandenberg, S. G. Assortative mating, or who marries whom? *Behavior Genetics* 2, 127–157 (1972).

2. Eaves, L. J. & Heath, A. C. Detection of the effects of asymmetric assortative mating. *Nature* 289, 205–206 (1981).

3. Buss, D. M. Human mate selection: opposites are sometimes said to attract, but in fact we are likely to marry someone who is similar to us in almost every variable. *American Scientist* 73, 47–51 (1985).

4. Heath, A. C. et al. No decline in assortative mating for educational level. *Behavior Genetics* 15, 349–369 (1985).

5. Heath, A. C. & Eaves, L. J. Resolving the effects of phenotype and social background on mate selection. *Behavior Genetics* 15, 15–30 (1985).

6. Heath, A. C., Eaves, L. J., Nance, W. E. & Corey, L. A. Social inequality and assortative mating: Cause or consequence? *Behavior Genetics* 17, 9–17 (1987).

7. Mare, R. D. Five decades of educational assortative mating. *American Sociological Review*. 56, 15–32 (1991).

8. Krueger, R. F., Moffitt, T. E., Caspi, A., Bleske, A. & Silva, P. A. Assortative mating for antisocial behavior: developmental and methodological implications. *Behavior Genetics* 28, 173–186 (1998).

9. Eaves, L. et al. Comparing the biological and cultural inheritance of personality and social attitudes in the Virginia 30,000 study of twins and their relatives. *Twin Research and Human Genetics* 2, 62–80 (1999).

10. Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R. J. & Rose, R. J. Assortative mating by body height and BMI: Finnish twins and their spouses. *American Journal of Human Biology* 15, 620–627 (2003).

11. Watson, D. et al. Match makers and deal breakers: analyses of assortative mating in newlywed couples. *Journal of Personality* 72, 1029–1068 (2004).

12. Conley D., Laidley T., Belsky D. W., Fletcher J. M., Boardman J. D., & Domingue B. W. Assortative mating and differential fertility by phenotype and genotype across the 20th century. *Proceedings of the National Academy of Sciences* 113, 6647–6652 (2016).

13. Kirkpatrick, M. Reinforcement and divergence under assortative mating. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267, 1649–1655 (2000).

14. Jiang, Y., Bolnick, D. I., & Kirkpatrick, M. Assortative mating in animals. *The American Naturalist* 181, E125–E138 (2013).

15. Stulp, G., Simons, M. J., Grasman, S., & Pollet, T. V. Assortative mating for human height: A meta-analysis. *American Journal of Human Biology* 29, e22917 (2017).

16. Eaves, L. J. Assortative mating and intelligence: An analysis of pedigree data. *Heredity* 30, 199–210 (1973).

17. Eaves, L. J. & Heath, A. C. Detection of the effects of asymmetric assortative mating. *Nature* 289, 205–206 (1981).

18. Maes, H., Neale, M. & Eaves, L. Genetic and environmental factors in relative body weight and human adiposity. *Behavior Genetics* 27, 325–351 (1997).

19. Zietsch, B. P., Verweij, K. J. H., Heath, A. C. & Martin, N. G. Variation in human mate choice: simultaneously investigating heritability, parental influence, sexual imprinting, and assortative mating. *The American Naturalist* 177, 605–616 (2011).

20. Keller, M. C. et al. The genetic correlation between height and IQ: Shared genes or assortative mating? *PLoS Genetics* 9, e1003451 (2013).

21. Robinson, M. R., et al. Genetic evidence of assortative mating in humans. *Nature Human Behaviour* 1, 1–13 (2017).

22. Landsteiner K. Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Zentralblatt für Bakteriologie, Parasitenkunde und Infektionskrankheiten* 27: 357–362 (1900).

23. Altham, P. M. The Measurement of Association of Rows and Columns for an r× s Contingency Table. *Journal of the Royal Statistical Society: Series B (Methodological)* 32, 63–73 (1970).

24. Long, J., & Ferrie, J. Intergenerational Occupational Mobility in Great Britain and the United States Since 1850. *The American Economic Review* 103, 1109–1137 (2013).

25. Eika, L., Mogstad, M., & Zafar, B. Educational Assortative Mating and Household Income Inequality. *Journal of Political Economy* 127, 2795–2835 (2019).

26. Reynolds, C. A., Barlow, T. & Pedersen, N. L. Alcohol, tobacco and caffeine use: spouse similarity processes. *Behavior Genetics* 36, 201 (2006).

27. Ask, H., Rognmo, K., Torvik, F. A., Røysamb, E. & Tambs, K. Non-random mating and convergence over time for alcohol consumption, smoking, and exercise: the Nord-Trøndelag Health Study. *Behavior Genetics* 42, 354–365 (2012).

28. Howe, L. J. et al. Genetic evidence for assortative mating on alcohol consumption in the UK Biobank. *Nature Communications* 10, 1–10 (2019).

29. Fox, G. A. Assortative mating and plant phenology: Evolutionary and practical consequences. *Evolutionary Ecology Research*, 5, 1–18 (2003).

30. van Doorn, G. S., Edelaar, P., & Weissing, F. J. On the origin of species by natural and sexual selection. *Science*, 326, 1704–1707 (2009).

31. Pusey, A., & Wolf M. Inbreeding avoidance in animals. *Trends in Ecology and Evolution*, 11, 201–206 (1996).

32. Falconer, D. S. Introduction to quantitative genetics (Ronald, 1960).

33. Nance, W. E., Corey, L. A., Rose, R. J. & Eaves, L. J. Relevance of the marriages of twins to the causal analysis of nonrandom mating. *Progress in Clinical and Biological Research* 69, 61–71 (1981).

34. Abdellaoui, A. et al. Association between autozygosity and major depression: stratification due to religious assortment. *Behavior Genetics* 43, 455–467 (2013).

35. Domingue B. W., Fletcher J. M., Conley D., Boardman J. D. Genetic and educational assortative mating among US adults. *Proceedings of the National Academy of Sciences* 111, 7996-8000 (2014).

36. Abdellaoui, A., Verweij, K. J. H. & Zietsch, B. P. No evidence for genetic assortative mating beyond that due to population stratification. *Proceedings of the National Academy of Sciences* 111, E4137 (2014).

37. Domingue B. W., Fletcher J. M., Conley D., Boardman J. D. Reply to Abdellaoui et al.: Interpreting GAM. *Proceedings of the National Academy of Sciences* 111, E4138 (2014).

38. Abdellaoui, A. et al. Educational attainment influences levels of homozygosity through migration and assortative mating. *PLoS One* 10, e0118935 (2015).

39. Tenesa, A., Rawlik, K., Navarro, P., & Canela-Xandri, O. Genetic determination of height-mediated mate choice. *Genome Biology* 16(1), 1–8 (2015).

40. Richardson, K., & Jones, M. C. Why genome-wide associations with cognitive ability measures are probably spurious. *New Ideas in Psychology* 55, 35–41 (2019).

41. Morris, T. T., Davies, N. M., Hemani, G., & Smith, G. D. Population phenomena inflate genetic associations of complex social traits. *Science Advances* 6, eaay0328 (2020).

42. Yengo, L., Sidari, M., Verweij, K. J., Visscher, P. M., Keller, M. C., & Zietsch, B. P. No evidence for social genetic effects or genetic similarity among friends beyond that due to population stratification: a reappraisal of Domingue et al (2018). *Behavior Genetics* 50, 67–71 (2020).

43. National Health and Family Planning Commission. The 2014 Statistical Communiqué on the Development of Health Services in China. http://www.nhc.gov.cn/guihuaxxs/s10742/201511/191ab1d8c5f240e8b2f5c81524e80f19.shtml. Retrieved 16 October 2022.

44. National Health and Family Planning Commission. The 2015 Statistical Communiqué on the Development of Health Services in China. http://www.nhc.gov.cn/guihuaxxs/s10748/201607/da7575d64fa04670b5f375c87b6229b0.shtml. Retrieved 16 October 2022.

45. Liu, J. et al. Seroepidemiology of hepatitis B virus infection in 2 million men aged 21–49 years in rural China: a population-based, cross-sectional study. *The Lancet Infectious Diseases* 16, 80-86 (2016).

46. Liu, J. et al. Frequencies and ethnic distribution of ABO and RhD blood types in China: a population-based cross-sectional study. *BMJ Open* 7, e018476 (2017).

47. Liu, J. et al. Maternal prepregnancy infection with hepatitis B virus and the risk of preterm birth: a population-based cohort study. *The Lancet Global Health* 5, e624-e632 (2017).

48. Hou, Y., Tang, K., Wang, J., Xie, D., & Zhang, H. Data from "Assortative Mating on Blood Type: Evidence from China" Tsinghua Cloud Storage. https://cloud.tsinghua.edu.cn/d/e86e227d8e66475ba790/. Deposited 2 September 2022.