Genetic and educational assortative mating among US adults

Benjamin W. Domingue[a,1], Jason Fletcher[b,c,d], Dalton Conley[e], and Jason D. Boardman[a,f]

[a]Institute of Behavioral Science and [f]Department of Sociology, University of Colorado Boulder, Boulder, CO 80309; [b]La Follette School of Public Affairs, [c]Center for Demography and Ecology, and [d]Department of Sociology, University of Wisconsin–Madison, Madison, WI 53706; and [e]Center for Genomics and Systems Biology, New York University, New York, NY 10003

Edited by Robert D. Mare, University of California, Los Angeles, CA, and approved April 16, 2014 (received for review November 15, 2013)

Understanding the social and biological mechanisms that lead to homogamy (similar individuals marrying one another) has been a long-standing issue across many fields of scientific inquiry. Using a nationally representative sample of non-Hispanic white US adults from the Health and Retirement Study and information from 1.7 million single-nucleotide polymorphisms, we compare genetic similarity among married couples to noncoupled pairs in the population. We provide evidence for genetic assortative mating in this population but the strength of this association is substantially smaller than the strength of educational assortative mating in the same sample. Furthermore, genetic similarity explains at most 10% of the assortative mating by education levels. Results are replicated using comparable data from the Framingham Heart Study.

homophily | random mating | genetic homogamy

Assortative mating occurs when individuals exhibit a preference for those who are either similar, (homogamy) or dissimilar (heterogamy) to themselves. Two expressions—"birds of a feather flock together" and "opposites attract"—are used to explain friendship and spousal pairings but denote opposite assumptions regarding the direction of selection. Critically, no existing research has quantified the degree to which individuals who select into a marriage are genetically similar to one another across the entire genome.

Quantifying genome-wide genetic assortative mating (GAM) in the population is important for methodological and substantive reasons. First, statistical models in genetic epidemiology, such as Hardy–Weinberg equilibrium, often assume random mating to forecast population allele frequencies, homozygosity rates, and other parameters of interest across generations (1) and behavior genetics models assume random mating to calculate heritability estimates (2). Second, social scientists have long studied the causes and consequences of assortative mating on a number of phenotypic measures such as height, education, religiosity, and political partisanship (3–5). Although there is research with a focus on the implications of genetic homogamy for phenotypic assortative mating (6), most studies of assortative mating have not considered the possibility that GAM may underlie phenotypic sorting. Social factors clearly limit opportunities to interact with people of different backgrounds (7, 8) but there is no study that simultaneously estimates educational assortative mating (EAM) and GAM in the population. Although much is known about changes in the nature of assortative mating over the past 50 y (5, 8, 9), little is known about the relationship between GAM and EAM.

We focus on EAM because it has received the largest amount of attention in the assortative mating literature (4) and, equally important, research has shown that educational attainment reflects genetic influences (10, 11). No existing study has used genome-wide data among spousal pairs to quantify GAM in the population. This observation coupled with the potential bias caused by GAM in traditional heritability estimates (12) makes this line of inquiry both substantively and methodologically important to a large group of biological and social scientists. In this paper we ask three related questions. First, is there any evidence of GAM in the population? Are genetically similar persons more likely to marry than genetically dissimilar persons both inclusive of and net of ethnic intramarriage? Or are spousal genotypes uncorrelated, as is sometimes assumed? Second, how does the magnitude of GAM compare with other phenotypically-based measures of assortative mating in the population—such as education? Third, to what extent is phenotypic assortative mating linked to GAM in the population?

Results

Estimates of EAM and GAM. EAM and GAM estimates from the Health and Retirement Study (HRS) (13) are shown graphically in Fig. 1. Fig. 1, *Upper* addresses the first two research questions in our study; *Upper Left* presents a graphical representation of GAM. To illustrate the meaning of this curve, consider the point where the two lines intersect. This point indicates that the median value of genetic similarity among spouses corresponds to the 55th percentile (the horizontal line) in the general population of all possible pairs; spouses are more genetically similar than randomly generated pairs in the population. To assess the magnitude of the increased spousal genetic similarity, we focus on the area of the shaded region above the 45° line. This produces an estimate of GAM of 0.045 [95% confidence interval (CI): 0.026, 0.061]. This estimate of GAM includes GAM due to intraethnic marriage among non-Hispanic whites, which we attempt to remove in subsequent analyses.

To gauge the magnitude of this GAM coefficient, we performed the same analysis using years of completed education plus a small amount of noise (the rationale for the inclusion of the noise is included in *SI Text, section S1*). This graph is shown in Fig. 1, *Upper Right*. Our estimate of EAM is 0.127 (95% CI: 0.109, 0.144), an estimate that is 2.9 times as large as our estimate

Significance

It is well established that individuals are more similar to their spouses than other individuals on important traits, such as education level. The genetic similarity, or lack thereof, between spouses is less well understood. We estimate the genome-wide genetic similarity of spouses and compare the magnitude of this value to a comparable measure of educational similarity. We find that spouses are more genetically similar than two individuals chosen at random but this similarity is at most one-third the magnitude of educational similarity. Furthermore, social sorting processes in the marriage market are largely independent of genetic dynamics of sexual selection.

Author contributions: B.W.D., J.F., D.C., and J.D.B. designed research; B.W.D. performed research; B.W.D. contributed new reagents/analytic tools; B.W.D. analyzed data; and B.W.D., J.F., D.C., and J.D.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: ben.domingue@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1321426111/-/DCSupplemental.

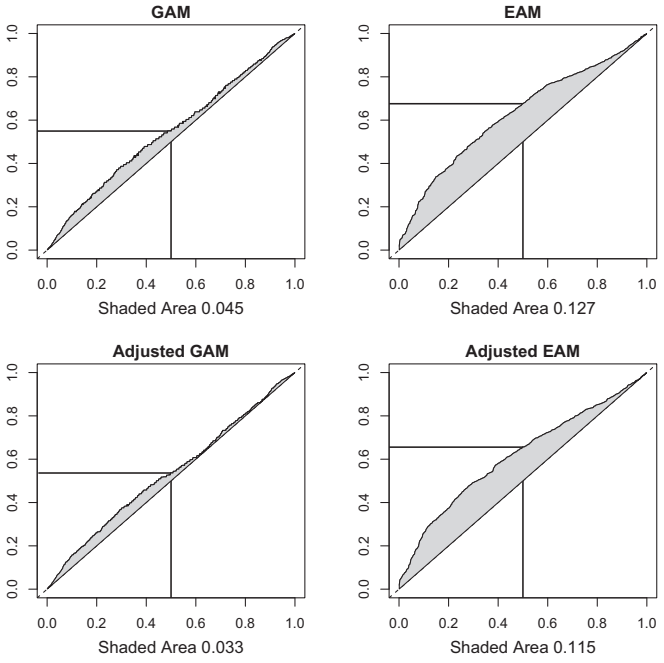

Fig. 1. Graphical representation of GAM and EAM. The *y* axis charts quantiles of the distribution of kinship or squared educational differences between all pairs. The *x* axis charts quantiles of the same distribution but restricted to just cross-sex white spousal pairs. The shaded area in each gives an estimate of assortative mating. The horizontal and vertical lines aid in interpretation. In *Upper Left*, one can observe that the genetic relatedness estimate at the 0.5 quantile of spousal pairs corresponds to the 0.55 quantile of all pairs. Adjusted GAM (*Lower Left*) includes control for same birth region (census division). Adjusted EAM (*Lower Right*) includes a control for kinship between the pairs.

of GAM. Together, these results answer our first two questions. Namely, GAM exists in this sample but it is substantially smaller in magnitude than EAM.

Next, we investigated whether GAM and EAM have a specific common explanation through a small set of SNPs related to educational attainment. We tested this hypothesis by examining proxies for the SNPs that reached genome-wide significance in a recent genome-wide association study (GWAS) on educational attainment (11). In particular, we conducted a χ^2 test using the sum of the risk alleles for the target SNPs for husbands and wives. The original SNPs were rs9320913, rs11584700, and rs4851266. We identified proxies using SNP Annotation and Proxy Search (14) that were correlated with the original SNP at no less than 0.8. The *P* values for the three tests were all above 0.35. Hence, we found little evidence that there was assortative mating based on these SNPs.

We conducted a replication analysis using data from the second generation of the Framingham Heart Study (15). It is important to note that the participants of this study are a group of predominantly white respondents from a geographically constrained area. In this secondary data set, we estimated GAM to be 0.025 (95% CI: 0.005, 0.046) based on 685 spousal pairs. Although we replicate the rejection of the null hypothesis of zero GAM in a second sample, we also note the decline in the magnitude of GAM compared with the estimate from HRS. Our estimated EAM in the Framingham sample was similar to the result from the HRS sample, 0.121 (95% CI: 0.102, 0.141).

Impact of Population Stratification on GAM. The existence of population stratification, small differences in allele frequencies that may exist across socially defined racial and ethnic groups, complicates many genetic analyses. In this section, we consider the extent to which population stratification may be present in our sample and how it may influence our measure of GAM. To characterize genetic divisions among the sample of non-Hispanic whites, we computed principal components (PCs) (*SI Text, section S2*) based on the complete set of SNPs. These methods consider the correlation between all of the SNPs within a population and identify factors that account for the greatest amount of common genetic variance. These factors align strongly with self-reported race and ethnicity and provide continuous measures of ancestry that are important controls for population stratification. There is substantial variability in the first PC only. Although we do not have information on ethnicity aside from Hispanicity, the PCs are largely unassociated with birth region (as a proxy for ethnic mixture). Differences in PCs may be capturing the genetic similarity (unrelated to population stratification) that we hope to investigate in our GAM analysis. As it is unclear if these PCs are confounding our estimate of GAM or are themselves an interesting component of GAM, we do not focus on estimates that control for these differences. We instead consider three alternative methods of adjusting for these differences in population stratification (estimates based on direct controls for PCs are shown in *SI Text, section S2*).

First, we use a subsample of our respondents with less variability in the first, and subsequent, PC(s) that presumably have less ethnic variability than the full sample. This should in turn reduce the impact that ethnic intramarriage among whites would have on our estimates. We estimated GAM among only those respondents with PC 1 > −0.003 to be 0.021 (95% CI: 0.002, 0.041). Note that this is very similar to the value obtained from our estimate of GAM in the Framingham Heart Study (15), a geographically homogenous sample.

A second approach for controlling the impact of population stratification is to control for birth region in our estimate of GAM because individuals from the same birth region are more likely to come from the same ethnic group than two individuals sampled from the entire nation (*SI Text, section S3* and ref. 16). In Fig. 1, *Lower Left*, we present an adjusted GAM estimate produced by residualizing kinship using a linear model with a dummy variable indicating whether a pair was born in the same census division. Based on this approach, we estimated an adjusted GAM of 0.033 (95% CI: 0.013, 0.049). This change suggests that some of our the initial GAM estimate is due to the fact that people from the same geographic area are more likely to marry one another than people from different areas (65% of the spousal pairs are from the same birth region compared with just 13% of nonspousal pairs) and that these geographic areas may capture subtle allele frequency differences across the population. That said, there is evidence for residual GAM, even with geographic controls. We note that this source of GAM is often not adjusted for in many estimates of heritability or demographic models of spousal assortative mating that use national (or international) samples, even if the samples are of non-Hispanic whites.

Finally, we also attempted to adjust for the influence of population stratification via direct manipulation of the genetic data. After computation of PCs, we identified SNPs that were most associated with the first five PCs (and thus potential ethnic markers) via GWAS. We then removed these SNPs from the genetic data and recalculated kinship (additional details on this process are in *SI Text, section S4*). Even after imposing extremely conservative restrictions that removed 70% of the SNPs (remaining SNPs were unrelated to any of the first five PCs), we estimated a GAM of 0.026 (95% CI: 0.005, 0.045). We discuss the relationship between the various estimates of GAM that control for population stratification in *Discussion*, but pause to note that several different approaches have converged on an estimate of residual GAM between 0.02 and 0.03.

Relationship Between GAM and EAM. To answer the third research question, we estimated EAM after first regressing out genetic similarity (based on the kinship estimates). Fig. 1, *Lower Right* describes the results of this analysis. As shown in this figure, adjusting for GAM reduced EAM to 0.115 (95% CI: 0.102, 0.133). Given that the kinship values used for this analysis may be affected by population stratification, we view this as an upper bound. Hence, at most 10% of the variance in EAM is due to GAM. We also examine this relationship in reverse by computing a GAM coefficient based on the residualized kinship coefficients (kinship was regressed on the squared educational differences of a pair). This coefficient declined from 0.045 to 0.026, a reduction of 42%. We discuss our interpretation of this result in the following section.

Discussion

Spouses are more genetically similar than two individuals chosen at random. As described in *SI Text, section S5*, our unadjusted GAM result of 0.045 suggests that a 1-SD increase in genetic similarity increases the probability of marriage by roughly 15%. This association is confounded, in part, by intraethnic marriage among whites but we continue to observe GAM even after a series of models designed to eliminate this source of assortative mating. That is, after replication with an independent dataset that is geographically homogeneous, restriction of our analyses to a genetically homogeneous subsample of respondents, adjustment of kinships for common birth region, and elimination from genetic data of SNPs that capture population structure, we obtain estimates of GAM between 0.02 and 0.03. The lack of additional ethnicity information in HRS makes it difficult to understand the quantity of GAM that is due to ethnic homogamy alone but the additional analyses suggest that preference for intraethnic marriage accounts for roughly one-half of observed GAM among non-Hispanic whites. It is worth noting that other phenomena could be related to both marriage preference and genetic architecture. Religion, for example, could be a source of GAM in this respect. Future research could consider the proportion of GAM that is due to such factors.

Although GAM exists, an important finding in our analyses is that the magnitude of GAM is significantly smaller than the magnitude of EAM. Furthermore, similar genotype explains only a small fraction of EAM (less than 10%). Our attempt to understand the amount of EAM that could be explained by GAM is based on the hypothesis that a fraction of phenotypic similarity is due to genetic similarity. In short, that GAM causes EAM. However, it is important for us to acknowledge that there are alternative explanations. Education could structure GAM through gene–environment correlations (17). For example, previous research (18) suggests that genetic similarity among friends is higher in schools with higher levels of economic inequality, which emphasizes the need to consider structural differences in educational institutions as a precursor to genetic selection into friendships. Our results (in particular, the 42% decline in GAM after controlling for EAM) indicate that social institutions may segregate people on genotype (presumably unwittingly), which could be behind some of the GAM that we observe. We do not assess this hypothesis empirically but we encourage others to consider this possibility in future research.

It is also important to note that both understandings (EAM causes GAM or GAM causes EAM) do not consider that this relationship is contingent upon the mean level of education among the pairs. For example, Eckland (19) hypothesizes that spousal correlations for intelligence are higher when the intelligence of either spouse is either exceptionally high or exceptionally low. This nonlinear relationship in conjunction with the strong correlation between intelligence and years of completed education suggests that the direction and magnitude of the GAM–EAM relationship may vary across the educational spectrum. Eckland (19) and others (20) have argued that assortative mating and the genetic influences on status-related outcomes may change over time. Higher levels of social inequality reduce the likelihood that otherwise small genetic factors will significantly shape an individual's socioeconomic attainment but historical changes in equality over time may provide or limit opportunities for these otherwise latent traits to manifest. Although it is unclear if the cohort range in the HRS is large enough to evaluate this hypothesis, we encourage future researchers to examine this possibility as well as the interactive hypothesis described above.

Our findings have important implications for a range of disciplines. Social scientists might gain additional understanding of assortative mating (or similar processes, such as friendship selection) by considering the role of genes. This is particularly important when one considers the significance of social factors that limit or enable two individuals to select into a relationship and how these factors differ across contexts and over time (18). Although it is beyond the scope of this paper, it is also important to consider the possibility that the intergenerational transmission of education may depend on the relative influence of EAM and GAM, which may change over time and context. That is, the influence of EAM on the intergenerational transmission of education may depend on the extent to which EAM is due to GAM. For example, if the proportion of EAM that is due to GAM is increasing over time, then it has important implications for our understanding of the intergenerational transmission of education. This perspective is not possible when one only examines EAM and offspring education.

Researchers presenting heritability estimates should consider including estimates of general assortative mating or trait specific genetic homogamy. Scientists have begun to interrogate the underlying assumptions of kinship based models that attempt to decompose the variation in a trait such as education into its additive genetic, common environmental, and unique environmental components. Recent work has used molecular approaches to test one major assumption: the equal environments assumption (21). The second key assumption, random mating with respect to the genetic architecture of the trait among the parental generation, has seen less investigation. Typically researchers use parental correlations in the phenotype as a rough estimate of nonrandom mating. However, of even greater value would be understanding the quantity of nonrandom mating that there is genetically with respect to the trait and how these associations have changed over time.

The results presented here only represent a first step in understanding the ways in which humans may assortatively mate with respect to their genome. For instance, an extensive literature (22) has emerged suggesting that heterosexual individuals find the odors of opposite sex persons more attractive if the test odor comes from someone who is genetically discordant on markers in the major histocompatiability complex area of chromosome six, which is thought to be under pressures of balancing selection. Such a region-specific, negative-assortative-mating dynamic may serve to depress overall (positive) GAM estimates. Thus, it may behoove future researchers to break apart the genome into parts that are relevant to specific pathways or processes that may be under different selective pressures to see if genome-wide GAM estimates mask a mixture of strong positive and negative dynamics with respect to different dimensions.

Our paper contributes to the literature on both GAM and EAM but has several limitations that we encourage others to consider. First, our results apply to opposite-sex non-Hispanic white pairs within the United States. For nonwhite pairs within the United States, different results might be obtained due to limited genetic variance among non-Hispanic whites compared with other groups (23) or because of different social contexts for non-Hispanic whites compared with others (e.g., the racial inequities that exist in the United States). That is, if individuals are selecting into a relationship because of genetic similarity,

then we might expect GAM to be higher among non-Hispanic whites who are less likely than others to face limitations in terms of residential, educational, or occupational choices. Second, patterns of GAM and EAM might differ in same-sex couples. Third, differences may be changing over time. For example, recent research (24) suggests that there has been a rise in assortative mating which has contributed to a rise in income inequality. Fourth, we estimated genetic similarity using SNPs from across the genome. Future research could focus on SNPs known to be important for education (11) or those identified in other GWAS to examine homogamy at a finer level than our whole-genome approach. Given our results from the SNPs implicated in the education GWAS, it might be that analyses at levels finer than the entire genome but much larger than a single SNP, such as chromosomes, would be appropriate.

Materials and Methods

Data. This paper uses data from the Health and Retirement Study (HRS) RAND fat files (13). Access to the genome-wide data was approved by National Center for Biotechnology Information Genotypes and Phenotypes Database (access no. 19335-3). Of the 9,429 individual with genetic data (described below), 4,584 were from the HRS cohort (five other cohorts are also included in the full data). Of the 4,584, there were 3,504 non-Hispanic whites. Of these, 1,763 individuals were in 862 spousal pairs (some individuals had more than one spouse). We focus on only those individuals (with complete data) in spousal pairs, 1,716 individuals in 825 spousal pairs, as there are differences between individuals in spousal pairs and those not in spousal pairs (e.g., spouses have roughly a quarter year of education more on average). These individuals were born during a large span of time (between 1920 and 1970) but the majority (59%) were born in the 1930s. To assess EAM, we used total years of education. In our sample, 14% had less than a high school education, 38% had a high school education, and the remainder had more than a high school education. We also used information on the respondent's birthplace (coded as one of nine census divisions plus two categories for US birth with no additional information and foreign birth, 0.1% and 5.1% of the sample, respectively).

Genetic data for the HRS is based on DNA samples collected in two phases. The first phase was collected via buccal swabs in 2006 using the Qiagen Autopure method. The second phase used saliva samples collected in 2008 and extracted with Oragene. Genotype calls were then made based on a clustering of both data sets using the Illumina HumanOmni2.5-4v1 array (details on the quality control process can be found via ref. 25). After standard quality control procedures (e.g., removing SNPs that were missing in more than 5% of samples; minor allele frequencies below 1%; failure to meet Hardy–Weinberg equilibrium, violations of which suggest errors in the genotyping process), we retained 1,707,214 SNPs. We also performed replication analysis on data from the Framingham Heart Study (15) (a description of these data can be found in *SI Text, section S6*).

Measuring Genetic Similarity. Quantifying GAM in the population relies on a valid and reliable measure of genetic relatedness between all individuals in the study. Genetic relatedness is a basic biological concept that undergirds quantitative genetic analyses (1). The bulk of this research relied on unmeasured genetic similarity among different types of relatives (e.g., siblings, twins, cousins, etc.) and recently this same conceptual approach used genome-wide data from related (26) and unrelated (27) individuals. These methods are similar in that they take advantage of naturally occurring variability in the degree to which two individuals' genomes are more or less similar compared with others in the population. It is precisely this variability between unrelated individuals that we use here. There are a number of methods for estimating genetic similarity based on measured genotype but the properties of these various estimates differ. We experimented with a measure that is based on the assumption of a common allele frequency across a sample (28) but this measure was found to be highly sensitive to population stratification (details are shown in *SI Text, section S7*). Therefore, we use a measure of kinship that has been shown to be more robust to population stratification than previous estimates of genetic similarity across the genome (29). This procedure produces a matrix that describes the genetic similarity for all pairs of individuals in our sample.

Measuring GAM. The traditional approach to measuring EAM is to analyze the correlation of spousal educational attainment. It is important to note that this approach is possible because each spouse has a level of education. In contrast, measures of genetic relatedness exist at the pair level because relatedness measures a quantity with respect to a specific alter, rather than an absolute level (e.g., years of completed schooling). Hence, a spousal pair would have only a single measure of genetic relatedness versus two measures of education, one for each spouse. The correlation approach is thus not a viable option for measuring GAM. We have instead chosen to concentrate on differences in the distributions of genetic relatedness between married and unmarried pairs of respondents. Although this approach is unique, we studied its behavior via a simulation study (*SI Text, section S5*), which demonstrated that the method is able to distinguish assortative mating from random mating in samples of this size.

Characterizing the presence and magnitude of genetic homogamy via a comparison of distributions is challenging because it requires a relevant comparison group. One approach would be to consider, for a focal individual, only those individuals with whom the individual is likely to marry given certain characteristics (e.g., age). Results based on such an approach would perhaps be unpersuasive given their potential sensitivity to the formation of the group of potential spouses for a person. To avoid this dilemma, we test GAM against the null hypothesis of random mating. As such, we make only minimal assumptions about the possible range of mates by restricting our comparisons of interest only to cross-sex, same-race individuals. We impose these sex and race restrictions due to limitations in existing data and methods. With respect to sex, we do not have data on same-sex couples. The restriction to same-race couples is done because the relatedness measures can be sensitive to population stratification that may exist across racial groups (additionally, there are relatively few cross-race couples in the data: only 6% of the spousal pairs from the 1,093 spousal pairs in the HRS cohort data discussed in *SI Text, section S7*).

For both EAM and GAM, our motivating counterfactual is that mates select at random into unions. As such, the distribution of educational or genetic differences among spousal pairs would be the same for all possible cross-sex and same-race pairs in the population. To test this assumption, we compute quantiles (0.001–0.999 in increments of 0.001) for the distribution of the differences among the spousal pairs. We then map these values among spousal pairs to the corresponding quantiles among nonspousal pairs (all cross-sex, same-race pairs). When such results are depicted graphically (Fig. 1), the 45° line indicates the null hypothesis that the similarity among spouses matches the similarity among nonspouses. If the similarity among spouses differs from the similarity of nonspouses, then this is captured by departure from the 45° line. EAM and GAM are estimated as the area between this curve and the 45° line. For key estimates, 95% CIs for the estimates were then created via 1,000 bootstrap replications.

When measuring EAM, we first standardize education within each sex. Our motivation for standardizing education with respect to sex is that more highly educated females will tend to marry more highly educated males. Because of the demographic composition of this cohort, "more education" might mean different things for males and females (e.g., "some college" for females versus a college degree for males). Without standardization, a monotonic relationship between the probability of marriage and educational differences cannot be assumed because there would be ambiguity about the region between 0 and the mean educational difference. That is, if the average difference in completed schooling between males and females is 2 y, a couple with the same level of schooling are not at the same point of their sex specific distribution of years of schooling, and are thus "different." For education, our results are comparable with and without standardization because the distributions across the genders are similar (*SI Text, section S1*). However, standardization is a potentially important component of the methodology and would be an important consideration if analyzing phenotypes, such as height, whose distributions vary more across sex. We also multiply all educational differences by −1 so that, as with kinships, larger numbers mean more similar respondents.

Population Stratification. Because racial/ethnic homogamy is already well known in the literature (30), we focus on residual GAM—GAM that remains within genetically stratified samples that may challenge the assumptions of random mating and intergenerational models in the social sciences. Thus, we only use a sample of non-Hispanic whites in the HRS. Intraethnic assortative mating among Americans of European descent is well documented (3) and small differences in allele frequencies across European ethnic groups are easily identified with genome-wide data (31). As such, the identification of GAM may simply show that Europeans with a similar ethnic background are more likely to marry one another than individuals from different ethnic backgrounds. For example, using data from the Framingham Heart Study, researchers decomposed total genetic variation into PCs that characterize these otherwise small genetic differences across European subpopulations and

they calculate a spousal correlation of 0.58 for the first PC in this sample (32). Using similar methods, we estimated a comparable value ($r = 0.54$) for the first PC among non-Hispanic and white spouses in the HRS. To identify residual GAM, we describe the results from a series of analyses that introduce restrictions in an attempt to understand the extent to which GAM may simply arise from ethnic homogamy within non-Hispanic white couples. These models include the following adjustments: (*i*) restriction of the sample based on the first PC, (*ii*) including statistical controls for census division of birth as a proxy for ethnic background, and (*iii*) estimating GAM with a reduced set of SNPs that do not show any evidence of stratification in our sample.

ACKNOWLEDGMENTS. This research uses data from the HRS, which is sponsored by the National Institute on Aging (Grants NIA U01AG009740, RC2AG036495, and RC4AG039029) and conducted by the University of Michigan. Research was supported by the Eunice Kennedy Shriver National Institute Of Child Health and Human Development (NICHD) of the National Institutes of Health (NIH) under Award R21HD078031. The authors also acknowledge cofunding from the NICHD and the Office of Behavioral and Social Sciences Research (1R21HD071884). Further support was provided by the NIH/NICHD-funded CU Population Center (R24HD066613). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1. Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics* (Pearson Education Limited, Essex, England), 4th Ed.
2. Neale MC, Cardon LR (1992) *Methodology for Genetic Studies of Twins and Families* (Kluwer Academic, Dordrecht, The Netherlands).
3. Schwartz CR (2013) Trends and variation in assortative mating: Causes and consequences. *Annu Rev Sociol* 39:451–470.
4. Blossfeld HP (2009) Educational assortative marriage in comparative perspective. *Annu Rev Sociol* 35:513–530.
5. Breen R, Salazar L (2011) Educational assortative mating and earnings inequality in the United States. *Am J Sociol* 117(3):808–843.
6. Thiessen D, Gregg B (1980) Human assortative mating and genetic equilibrium: An evolutionary perspective. *Ethol Sociobiol* 1:111–140.
7. Blau P (1994) *Structural Contexts of Opportunities* (Univ of Chicago Press, Chicago).
8. Mare RD (1991) Five decades of educational assortative mating. *Am Sociol Rev* 56(1): 15–32.
9. Schwartz CR, Mare RD (2005) Trends in educational assortative marriage from 1940 to 2003. *Demography* 42(4):621–646.
10. Branigan AR, McCallum KJ, Freese J (2013) Variation in the heritability of educational attainment: An international meta-analysis. *Soc Forces* 92(1):109–140.
11. Rietveld CA, et al.; LifeLines Cohort Study (2013) GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 340(6139): 1467–1471.
12. Plomin R, DeFries JC, Roberts MK (1977) Assortative mating by unwed biological parents of adopted children. *Science* 196(4288):449–450.
13. Center for the Study of Aging (2014) RAND Enhanced Fat Files (RAND Corporation, Santa Monica, CA). Available at www.rand.org/labor/aging/dataprod/enhanced-fat.html. Accessed April 29, 2014.
14. Johnson AD, et al. (2008) SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24(24):2938–2939.
15. National Center for Biotechnology Information dbGaP (2007) Framingham SNP Health Association Resource (National Center for Biotechnology Information, Bethesda, MD). Available at www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v1.p1. Accessed April 29, 2014.
16. Lieberson S, Waters MC (1988) *From Many Strands: Ethnic and Racial Groups in Contemporary America* (Russell Sage Foundation, New York).
17. Shanahan MJ, Hofer SM (2005) Social context in gene-environment interactions: Retrospect and prospect. *J Gerontol B Psychol Sci Soc Sci* 60(Spec No 1, Special Issue):65–76.
18. Boardman JD, Domingue BW, Fletcher JM (2012) How social and genetic factors predict friendship networks. *Proc Natl Acad Sci USA* 109(43):17377–17381.
19. Eckland BK (1972) Evolutionary consequences of differential fertility and assortative mating in man. *Evol Biol* 5:293–305.
20. Adkins DE, Guo G (2008) Societal development and the shifting influence of the genome on status attainment. *Res Soc Stratif Mobil* 26:235–255.
21. Conley D, Rauscher E, Dawes C, Magnusson PKE, Siegal ML (2013) Heritability and the equal environments assumption: Evidence from multiple samples of misclassified twins. *Behav Genet* 43(5):415–426.
22. Milinski M (2006) The major histocompatibility complex, sexual selection, and mate choice. *Annu Rev Ecol Evol Syst* 37:159–186.
23. Tishkoff SA, Gonder MK (2007) *Human Origins Within and Out of Africa* (Cambridge Univ Press, Cambridge, UK).
24. Greenwood J, Guner N, Kocharkov G, Santos C (2014) *Marry Your Like: Assortative Mating and Income Inequality* (National Bureau of Economic Research, Cambridge, MA), NBER Working Paper No. 19829.
25. National Center for Biotechnology Information dbGaP (2012) Health and Retirement Study (National Center for Biotechnology Information, Bethesda, MD). Available at www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000428.v1.p1. Accessed April 29, 2014.
26. Visscher PM, et al. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* 2(3):e41.
27. Yang J, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565–569.
28. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* 88(1):76–82.
29. Manichaikul A, et al. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867–2873.
30. Lin KH, Lundquist J (2013) Mate selection in cyberspace: The intersection of race, gender, and education. *Am J Sociol* 119:183–215.
31. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456(7218): 98–101.
32. Sebro R, Hoffman TJ, Lange C, Rogus JJ, Risch NJ (2010) Testing for non-random mating: Evidence for ancestry-related assortative mating in the Framingham heart study. *Genet Epidemiol* 34(7):674–679.

Supporting Information

Domingue et al. 10.1073/pnas.1321426111

SI Text

S1. Sensitivity of Educational Assortative Mating Estimate

We address two issues in the computation of educational assortative mating (EAM): the addition of small quantities of noise and the within-sex standardization of educational attainment. Working with the raw squared educational differences leads to inaccurate results for EAM. The reason for this is subtle. The left-hand side of the distribution of educational differences is a long string of 0s (those pairs with the same education). Any quantile, no matter how small, computed relative to this empirical cumulative distribution function is going to be the percentage of pairs with the same education. Because this is a rather sizeable percentage of the overall distribution when there is no measurement error, the area between the curves is distorted. For this reason, we instead worked with education that was slightly perturbed at the individual level by adding a very small amount of noise. To demonstrate the robustness of our finding to this addition of noise, we conducted a sensitivity analysis in which the SD of the noise varied. Results are shown in Fig. S1. When the distribution of noise is quite large (SD = 1), the signed area starts at around 0.11. As the SD decreases to very near 0, the signed area settles around the estimate from Fig. 1. The far right-hand dot in Fig. S1 represents the signed area when no error is used. We also considered an estimate of EAM in which education was not standardized. The resulting EAM estimate, 0.131, was quite similar to the estimate 0.127 presented in *Results*, *Estimates of EAM and GAM*. This lack of a change is due in part to the fact that the educational differences between males and females in our sample were fairly small (a median of 12 y for both genders and only a difference of 0.2 y in the means). These results provide confidence that our approach for EAM measurement is not a remnant of modeling decisions.

S2. Principal Components

Fig. S2 shows the first four principal components (PCs) for the sample of spouses. These PCs were computed within the non-Hispanic white sample of respondents that are analyzed in Fig. 1. There is substantially more variation on the first PC than on any of the others. There is no information on ethnicity aside from Hispanicity in the Health and Retirement Study (HRS) (1), so we used the region of birth as one way of characterizing the PCs. Fig. S3 shows the mean by census division for PCs 1 and 2. The scale of this figure is based on the range of the individual values of the PCs (and the vertical line represents a cutoff to be discussed shortly). In brief, the PCs did not sharply distinguish between regions although one can see that the Atlantic seaboard (regions 1 and 2) tended to have slightly lower values on PC 1 than the other regions.

Analyses in which the kinships were adjusted for pairwise difference in either the squared or absolute value of the PCs are described in Table S1. After adjusting for just PC 1, the genetic assortative mating (GAM) estimates declined substantially. Adjusting for additional PCs moved the estimates to nearly 0. For example, adjusting for the first PC reduces GAM to 0.011 [95% confidence interval: −0.006, 0.029]. As described in *Impact of Population Stratification on GAM*, we believe that this approach is potentially flawed because it is unclear what differences between individuals (geographic differences? differences in countries of origin?) are being captured by the PCs. Turning back to Fig. S2, the red dots were chosen as a subset of the spousal sample (PC 1 > −0.003) that was relatively comparable on these PCs. We estimated a GAM value of 0.021 among this sample. This estimate is comparable to the value found among the ethnically homogenous Framingham sample described in *Description of Framingham Data*.

S3. Geography as a Proxy for Ethnicity

In this section, we present evidence that controls for the census division capture regional variability in ethnicity. The census divisions partition the states in the following way:

1) New England division: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont
2) Middle Atlantic division: New Jersey, New York, and Pennsylvania
3) East North Central division: Illinois, Indiana, Michigan, Ohio, and Wisconsin
4) West North Central division: Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota
5) South Atlantic division: Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, and West Virginia
6) East South Central division: Alabama, Kentucky, Mississippi, and Tennessee
7) West South Central division: Arkansas, Louisiana, Oklahoma, and Texas
8) Mountain division: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming
9) Pacific division: Alaska, California, Hawaii, Oregon, and Washington

HRS contains data on the census division at birth for each respondent (unspecified US births and foreign births are coded as 10 and 11, respectively; see Fig. S3). We used data from the 1980 US census (2) to compare ancestry within and across census divisions. We focused on spouses living in the same households, with valid ancestry records, and born between 1930 and 1940 (to make the respondents comparable to the sample of HRS respondents used here). After imposing these filters, we had 650,724 individuals of European ancestry and 165,552 individuals of non-European ancestry. We define European ancestry based on the ANCESTR1 variable, specifically codes 1–195. These codes correspond to numerous countries or regions across Western and Eastern Europe. However, we have excluded those of European Hispanic origin to be consistent with the exclusion of Hispanics from the HRS dataset.

To determine ethnic concentration within census divisions, we computed the mean percentage of individuals identifying as a particular ancestry within a census division. A lower value indicates a more diverse set of ancestries within a region. Suppose one state was evenly split between white, black, and Hispanic individuals. A second state was evenly split between white and Asian individuals. The index for the first state would be one-third whereas the index for the second state would be one-half. The first state, with the lower value, has the more diverse population. We then divided this index by the mean across the entire nation. We define this as the ethnic concentration within a region. Within a census division, the average concentration of Europeans European was 1.6. Within states, the average concentration was 2.9. Clearly there is more ethnic concentration within states, but census divisions explain a proportion of this. It is interesting to note that there is much greater concentration of ethnicities within both census divisions (2.5) and states (6.5) when non-European ethnicities are considered.

We can also use this data to understand the tendency toward intraethnic marriages and the relationship between intraethnic marriages and place of birth. Among 195,355 spousal pairs (where

each spouse is of European ancestry), 41% of the pairs were of the same ancestry. To interpret this number, the fact that the ancestry indicator is relatively fine-grained (over 100 different ancestry designations) must be remembered. We also considered the following hierarchical regression model:

$$\text{logit}\left(\Pr\left(\text{Same\_Ancestry}_{ijk}=1\right)\right)=\alpha+\beta\cdot\left(\text{Same\_Division}_{ijk}\right)+\mu_j+\gamma_k$$

for pair i in census division j and state k. Being born in the same census division increases the odds of a marriage between individuals of the same ancestry by 70%. The variance components associated with μ_j and γ_k were 0.11 and 0.06, respectively. Based on this evidence, we argue that there is clearly ethnic concentration among individuals of European ancestry in the United States that is captured by geography. Furthermore, we argue that being born in the same census division explains some of the preference for intraethnic marriages in the United States.

S4. Removal of SNPs Associated with PCs

We identified SNPs associated with population stratification by performing a genome-wide association for each of the first five PCs (controlling for sex and birth year). We then systematically removed those SNPs from our genetic data which had a P value in one of the five regressions that was below a given threshold (this varied from 5e-8 to 5e-2). With these different sets of SNPs, we then recomputed kinship values based on the remaining SNPs and reestimated GAM and adjusted GAM (based on controlling for census division of birth). The results of this exercise are presented in Table S2. Note that we lose over 70% of the SNPs going from the full genetic sample to only those SNPs with P values from all five regressions greater than 0.05. These remaining 457,201 SNPs are those that show very little evidence of population stratification in our sample. The most important observation is that our estimated GAM is relatively insensitive to the removal of SNPs until we get to the 5e-3 threshold, where nearly half of the SNPs have been removed. However, even after the removal of the majority of the SNPs, there is still evidence for GAM. Furthermore, the reduction due to the adjustment (based on same census division at birth) is much less for these estimates based on kinship computed using only SNPs unassociated with the first five PCs.

S5. Simulation Study

The proposed methodology is, to our knowledge, unique in the study of homogamy. Hence, it is important to determine that it is a viable approach for detecting homogamy in our sample. This simulation study demonstrates two crucial facts. First, the methodology can distinguish assortative mating from random mating. Second, the results produced by the methodology vary as expected as a function of the strength of assortative mating. The simulation study presented here is based on systematically controlling the strength of homogamy in a simulated sample and then calculating the area (as described in *Materials and Methods*), which acts as a measure of assortative mating.

The simulation involves three key steps. In the below description of the simulation, it is important to remember that there are in fact two simulation studies (one for kinship, one for educational differences) that share a common structure. For fixed values of the sample size (N), homogamy strength (indexed by A, described below), and SD of kinship values (σ^2), consider the following:

i) For each pair of individuals, a quantity is randomly generated that represents either genetic relatedness or the squared difference in years of education. Consider first relatedness. We simulate relatedness values by sampling $(N^2-N)/2$ (this is the number of lower-triangular entries in an $N\times N$ matrix) values from Normal[0, σ^2]. We use the observed SD for kinships in our sample as the value of σ^2. For education, we first generate individual-level educations using the observed distribution of educations in our sample and then generate all possible squared pairwise differences.

ii) We now let individuals select into unions. Individuals select into pairs based on a multinomial distribution. The procedure differs for education and kinship. Consider the set of relatedness estimates for all individuals with individual *i*. If individuals k and i have relatedness R_{ik}, then a weight (proportion to the probability of individual k marrying individual i) is assigned to individual k:

$$w_k=\frac{\exp(AR_{ik})}{1+\exp(AR_{ik})}.$$

The degree of homogamy in the simulation is manipulated through A. When $A = 0$, there is no homogamy (mating is random with respect to relatedness) and this is reflected by all pairings getting equal weights. The weights are then standardized to sum to unity and are the probabilities for the multinomial distribution. A draw from multinomial distribution (with only a single trial) is used to generate a mate for individual i. Mates are generated for all individuals in this manner with the additional restriction that only a single mate is assigned to each person.

To understand the computation of the weights for education, it is important to be aware of a key distinction between kinship and education. With kinship, we have more and less related individuals and there should be a monotonically increasing relationship between relatedness and the probability of getting married. This is the motivation behind the choice of the logistic transformation above. With squared education differences, not only is the distribution bounded below by 0, but the relationship should also be monotonically decreasing (increasing differences in education should lead to decreasing probabilities of getting married). This requires a different transformation and we use

$$w_k=\frac{1}{1+AD_{ik}},$$

where D_{ik} is the squared education difference between two individuals. Again, A is used to control the strength of homogamy in the simulation. Once weights are computed, the same procedure is used to generate matched pairs.

iii) The signed area metric is then computed based on the distribution of spousal differences to differences between all pairs. Unlike in the main text, we do not multiply educational differences by −1. This is done to emphasize the difference between educational differences and genetic relatedness values in the simulation.

The simulation performs those steps for a fixed value of N (chosen to replicate the number of spousal pairs, $n = 825$, in our sample) and different values of A.

Key results for the simulation study are shown in Fig. S4. The y axis in this figure is the signed area as described in *Materials and Methods*. The x axis measures changes in A that are being manipulated in the simulation. This quantity controls the probability weight of two individuals marrying. The scale factor is based on computations involving the distribution of either pair relatedness or educational differences. In particular, it is the value of the ratio of the probability weight at one SD above the mean of this distribution to the value at the mean. When this value is unity, there is no assortative mating (e.g., the random mating hypothesis is true). Note that in both the kinship and education versions of the simulation (gray and black) a value of

unity corresponds to essentially no area between the assortative mating curve and the 45° line. This indicates that the methodology can identify situations in which mating is random. Furthermore, as the scale factor deviates from unity we are able to detect increasing GAM and EAM (signed areas deviate from 0).

The estimated EAM coefficient from Fig. S4 (dashed black line) is consistent with a scale factor of roughly 0.5. This indicates that, according to the assumptions in our calculation, a 1-SD increase (from the mean) in squared educational differences corresponds to a probability weight that is 50% as large as the one used at the mean squared education difference. The result for GAM (dashed gray line) is weaker. A 1-SD increase from the mean relatedness corresponds, under the assumptions of our simulation, to a 15% increase in the probability weight of two individuals marrying. These results provide intuition regarding the probability of marriage that is consistent with the homophily observed via the signed areas. However, they must be interpreted with care because they are dependent on the assumptions used in this simulation.

S6. Description of Framingham Data

The study sample for this project was derived from the Framingham SNP Health Association Resource (SHARe, Version 6) as available through the National Center for Biotechnology Information Database of Phenotypes and Genotypes dbGaP (3). The original cohort of the study was first assessed in 1948; nearly 25 y later, the respondents' children (the G2 sample, $n = 3{,}548$) and many of their spouses participated in this study of the offspring cohort. Then, in 2002, roughly 4,000 adults who had at least one parent in the offspring cohort took part in the third generation (G3) cohort study. The analysis for our research focused on 1,624 individuals from the G2 sample of the Framingham Heart Study. We use genetic data for 260,402 SNPs, details on the quality control process used to select these SNPs can be found elsewhere (4). Using 685 spousal pairs drawn from the sample of 1,624 individuals, we calculated a GAM value of 0.025 (0.005, 0.046).

S7. Sensitivity of Genome-Wide Complex Trait Analysis Estimates to Population Stratification

We considered two measures of genetic relatedness. The first (5) assumes a common allele frequency. These estimates, which we refer to as "GCTA" (genome-wide complex trait analysis) estimates based on the software used to generate them, are likely to be biased in the presence of population stratification. To demonstrate this, we computed GCTA estimates as well as kinship estimates (6) for all spouses in the HRS cohort (2,163 individuals in 1,093 spousal pairs). Of the individuals in this sample, 8% identified as black/African-American and 3% identified as other. Fig. S5 compares GCTA and kinship estimates for the two individuals. Note the large estimates for couples that consist of two black individuals (represented as "5" in the figure). In contrast, pairs that consist of white couples have reasonable GCTA estimates; the interquartile range was between 0.01 and 0.015. Because the GCTA estimates were based on a primarily white sample, we inaccurately conclude that two nonwhite individuals are genetically quite similar. The median black spousal pair has a GCTA estimate approach of 0.39. This approaches the estimated genetic relatedness of full siblings from other studies (especially figure 1 of ref. 7). This is clearly a problem. In contrast, the kinship estimates never exceed 0.05, which is to be expected given that these are unrelated people. Some pairs have large negative values, but these are typically between couples with different racial backgrounds and would thus be excluded from our analyses.

1. Center for the Study of Aging (2014) RAND Enhanced Fat Files (RAND Corporation, Santa Monica, CA). Available at www.rand.org/labor/aging/dataprod/enhanced-fat.html. Accessed April 29, 2014.
2. Ruggles S, et al. (2010) Integrated Public Use Microdata Series: Version 5.0 (Univ of Minnesota, Minneapolis) [Machine-readable database].
3. National Center for Biotechnology Information dbGaP (2007) Framingham SNP Health Association Resource (National Center for Biotechnology Information, Bethesda, MD). Available at www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v1.p1. Accessed April 29, 2014.
4. Boardman JD, et al. (2013) Is the Gene-Environment interaction paradigm relevant to Genome-Wide studies? The case of education and body mass index. *Demography* 51(1):119–139.
5. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* 88(1):76–82.
6. Manichaikul A, et al. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867–2873.
7. Visscher PM, et al. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* 2(3):e41.

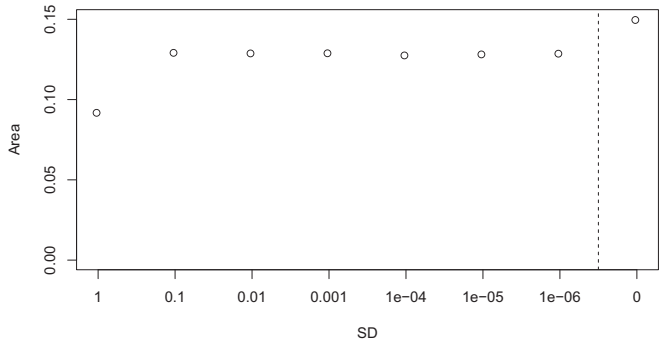

Fig. S1. EAM estimate as a function of the SD of the added noise.

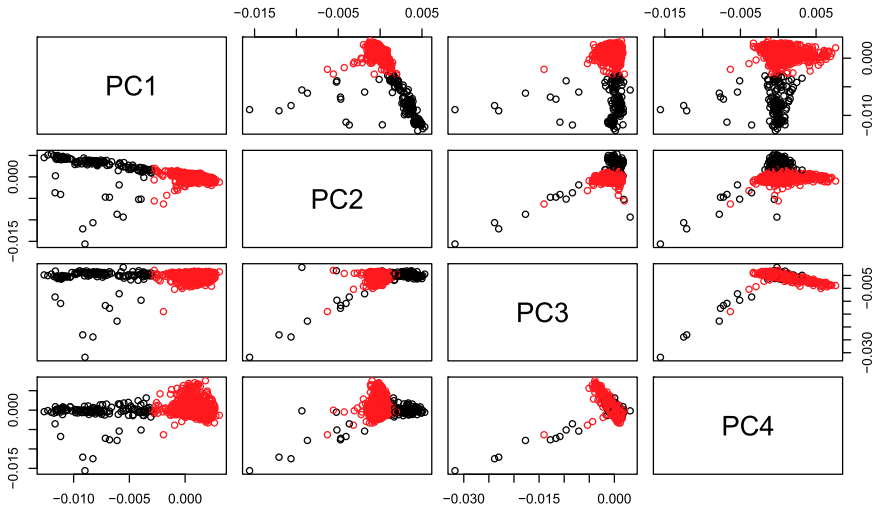

Fig. S2. Matrix scatterplot of the first four genome-wide PC values. All data comes from the HRS (n = 1,716) (1). Red dots identify respondents with PC 1 > −0.003. This is a more genetically homogeneous subgroup that is the focus of additional analyses.

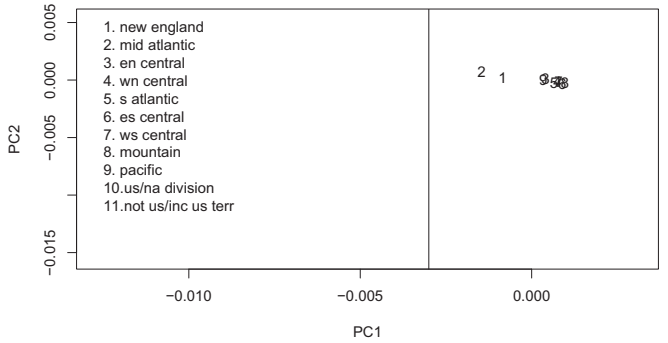

Fig. S3. Geographic clustering in the first two PCs. All data comes from HRS (1). The vertical line indicates the threshold for identification of the homogeneous subgroup (the red dots in Fig. S2). Divisions 3–8 are largely overlapping showing genetic similarity, at least with respect to the first two PCs, in these census divisions.

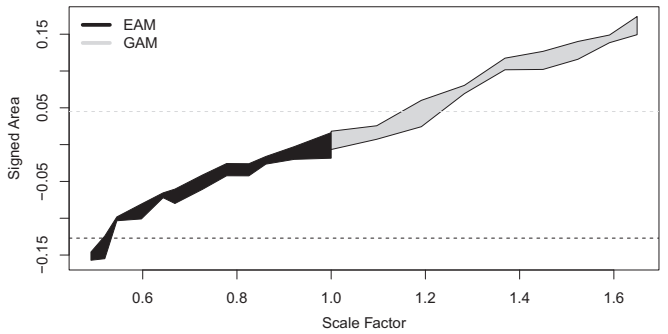

Fig. S4. Simulation results: Polygons show 10th and 90th percentile of the homogamy estimates for 150 iterations at specified values of the scale factor.

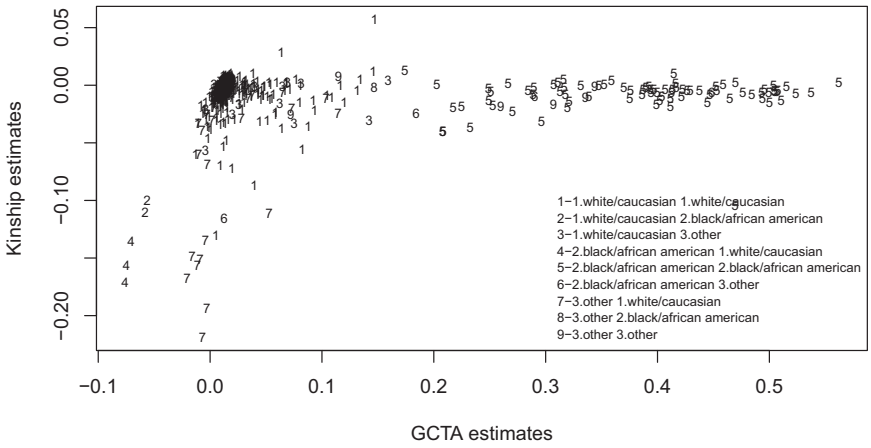

Fig. S5. Comparison of GCTA and kinship estimates of genetic similarity in the HRS (1).

Table S1. GAM estimates after controlling for PCs, both squared differences and absolute values

Controls	Squared differences	Absolute values
PC 1	0.011	0.008
PCs 1–2	0.012	0.008
PCs 1–3	0.013	0.005
PCs 1–4	0.009	0.006
PCs 1–5	0.009	0.006

Table S2. Results obtained after removal of SNPs associated with population stratification

Threshold	No. of SNPs	GAM	Adjusted GAM
None	1,707,214	0.045	0.033
5E-08	1,516,889	0.043	0.030
5E-06	1,431,983	0.043	0.029
5E-04	1,201,518	0.040	0.030
5E-03	934,430	0.036	0.026
5E-02	457,201	0.026	0.020