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.

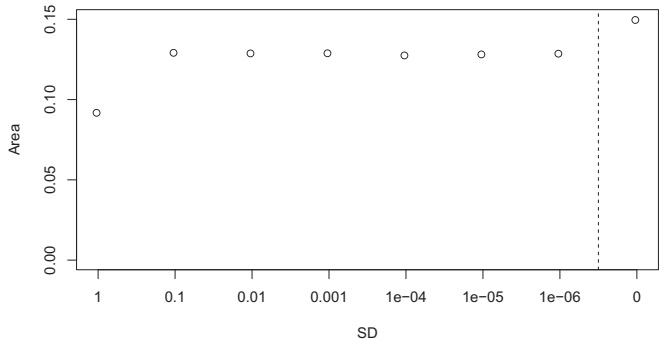

Fig. S1. EAM estimate as a function of the SD of the added noise.

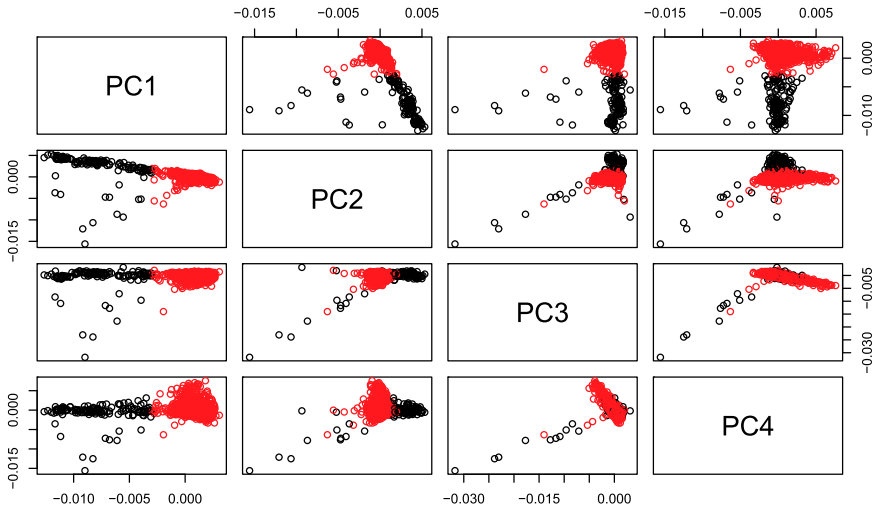

Fig. S2. Matrix scatterplot of the first four genome-wide PC values. All data comes from the HRS (n = 1,716) (1). Red dots identify respondents with PC 1 > −0.003. This is a more genetically homogeneous subgroup that is the focus of additional analyses.

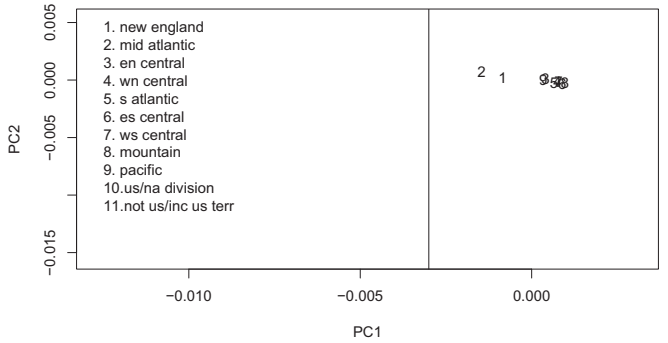

Fig. S3. Geographic clustering in the first two PCs. All data comes from HRS (1). The vertical line indicates the threshold for identification of the homogeneous subgroup (the red dots in Fig. S2). Divisions 3–8 are largely overlapping showing genetic similarity, at least with respect to the first two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

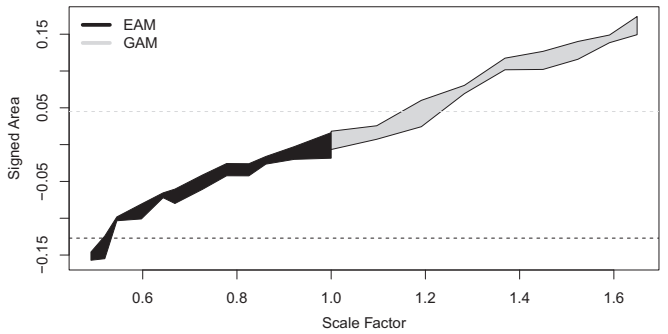

Fig. S4. Simulation results: Polygons show 10th and 90th percentile of the homogamy estimates for 150 iterations at specified values of the scale factor.

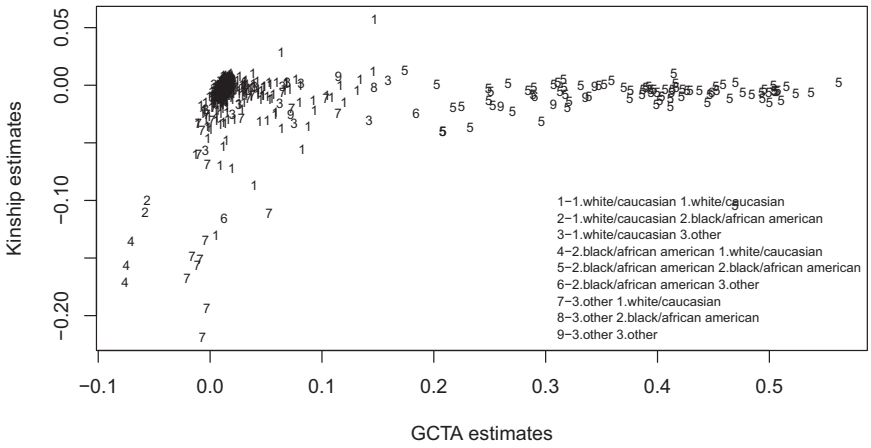

Fig. S5. Comparison of GCTA and kinship estimates of genetic similarity in the HRS (1).

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