Demography (2023) 60(6):1631–1648 Published online: 8 November 2023
DOI 10.1215/00703370-11053145

Understanding Internal Migration: A Research Note Providing an Assessment of Migration Selection With Genetic Data

Shiro Furuya, Jihua Liu, Zhongxuan Sun, Qiongshi Lu, and Jason M. Fletcher

ABSTRACT Migration is selective, resulting in inequalities between migrants and nonmigrants. However, investigating migration selection is empirically challenging because combined pre- and post-migration data are rarely available. We propose an alternative approach to assessing internal migration selection by integrating genetic data, enabling an investigation of migration selection with cross-sectional data collected post-migration. Using data from the UK Biobank, we utilized standard tools from statistical genetics to conduct a genome-wide association study (GWAS) for migration distance. We then calculated genetic correlations to compare GWAS results for migration with those for other characteristics. Given that individual genetics are determined at conception, these analyses allow a unique exploration of the association between pre-migration characteristics and migration. Results are generally consistent with the healthy migrant literature: genetics correlated with longer migration distance are associated with higher socioeconomic status and better health. We also extended the analysis to 53 traits and found novel correlations between migration and several physical health, mental health, personality, and sociodemographic traits.

KEYWORDS Migration • Biodemography • Genome-wide association study • UK Biobank

Introduction

Multiple theories posit that migrants are not randomly selected from a population. Examples include the healthy migrant hypothesis (Jasso et al. 2004; Palloni and Arias 2004; Palloni and Morenoff 2006), Borjas' (1987) application of the Roy (1951) model of selection in the economics literature to migration, and Ravenstein's (1885) law of migration. One illustrative migration selection process is that skilled and healthy individuals are more likely to migrate than less skilled and unhealthy counterparts because these qualities are necessary for the benefits of migration to outweigh its economic, personal, physical, and psychological costs (see Feliciano 2020).

Much research has presented empirical evidence supporting this positive migration selection. For example, immigrants in the United States, particularly long-distance

ELECTRONIC SUPPLEMENTARY MATERIAL The online version of this article (https://doi.org/10.1215/00703370-11053145) contains supplementary material.

migrants, tend to be more educated than those remaining in their home countries (Feliciano 2005). Earlier research also found better health among U.S. immigrants relative to nonmigrants residing in countries of origin, consistent with positive health selectivity of migration (Bostean 2013; Crimmins et al. 2005; Morey et al. 2020; Ro et al. 2016; Rubalcava et al. 2008).[1] Similarly, research in Europe has demonstrated that migrants have higher childhood socioeconomic status (SES) and health than nonmigrants in their sending countries (Fuller-Thomson et al. 2015; Schmidt et al. 2022), suggesting positive migration selection. Importantly, migration selection on SES and health can be found in internal migration contexts (Borjas et al. 1992; Lu 2008; Nauman et al. 2015; Rauscher and Oh 2021; Wilding et al. 2016).[2] Overall, international and internal migration are highly selective along many dimensions of SES and health.

Despite these established theoretical frameworks, data availability is a crucial limitation for studies examining migration selection. Given the effects of migration on migrants' SES and health (Lu 2010), a simple comparison of SES and health between migrants and nonmigrants reflects both selection and causation of migration. Innovative research has examined migration selection using longitudinal data that include both pre- and post-migration information (Abramitzky et al. 2012; Fuller-Thomson et al. 2015; Lu 2008; Nauman et al. 2015; Rubalcava et al. 2008), but such data are rarely available. This data constraint generally prevents scholars from separating migration selection and migration effects (Darlington et al. 2015).

The issue of data availability in migration studies goes beyond the lack of longitudinal data tracking migration behaviors. Prior research examining the healthy migrant hypothesis relied on subjective health assessments (Akresh and Frank 2008; Mehta and Elo 2012; Nauman et al. 2015).[3] However, these measures might partially reflect systematic differences in reporting tendencies between sociodemographic groups (Altman et al. 2016; Grol-Prokopczyk et al. 2011; Rossouw et al. 2018). An alternative to self-assessment is biomarker data, which can represent objective measures of risks of future diseases (Crimmins et al. 2010; Harris and Schorpp 2018). Thus, biomarker measures might uncover migration selection in latent health risks that do not appear in subjective health assessments. This feature of biomarkers is important in a case such as internal migration in the United Kingdom, where migration is concentrated among young adults (Bernard et al. 2016), who are less likely to perceive health issues.

We propose a novel approach to assess migration selection using a combination of standard genomic analysis toolkits: a genome-wide association study (GWAS) and genetic correlation analysis.[4] We first explore genetic variants correlated with

[1] Some of these studies showed negative migration selection on self-reported health (Bostean 2013; Rubalcava et al. 2008).

[2] For example, in the United States between 1880 and 1990, Black migrants from the South to the North had higher educational attainment than Black nonmigrants in the South (Tolnay 1998).

[3] Some studies used biomarkers (Beltrán-Sánchez et al. 2016; Crimmins et al. 2005; Riosmena et al. 2013; Rubalcava et al. 2008), but their biomarker variation was limited.

[4] A GWAS is a hypothesis-free scan of the genome that estimates statistical associations between each genetic location (variant) and an outcome of interest. Estimates from a GWAS can then be used in several types of downstream analysis. Genetic correlation analysis compares the similarity of GWAS estimates for one outcome (in this case, migration) with GWAS estimates from other outcomes (here, SES and health) to assess an overall genetic correlation among the outcomes. Alternatively, GWAS estimates can

migration through a GWAS and then use a genetic correlation analysis to assess whether and how these genetic variants correlated with migration are also associated with SES and health. Given that skilled and healthy individuals are selected to migrate, migrants are expected to have genetic traits correlated with higher SES and better health than nonmigrants: genetic variants correlated with migration will also be correlated with higher SES and better health.

The framework takes advantage of the fact that genetic variants are determined at conception, remain unchanged throughout the life course, and thus cannot be affected by migration, SES, or self-assessed health.[5] These qualities allow us to rule out migration effects and collect these measures post-migration. The broad scope of prior genetic analysis also enables us to consider many traits in our genetic correlation analysis, even those that affect older individuals relative to our sample. Together, these methods allow novel correlations between migration and other traits that are not typically measured (or cannot be measured). Additionally, future research can use our GWAS findings for migration in downstream analysis of migration in smaller datasets that contain genetic data, such as the Health and Retirement Study and the English Longitudinal Study of Aging. Furthermore, our results will directly show the role of genetics in migration selection, which demographers have suggested (Palloni and Arias 2004). Overall, our exploration of genetic correlations between migration and SES and health sheds light on understudied but potentially important dimensions of migration selection and integrates a social genomics approach into the migration literature.

Data and Methods

The UK Biobank (UKB) is a large-scale biobank study of more than 500,000 people that collected baseline data in 2006–2010. The UKB recruited the baseline sample through an invitation letter sent to individuals aged 40–69 who were living reasonably close to one of the 22 catchment areas where UKB assessment centers were located (see Figure 1). The UKB is suitable for our purposes because it includes a large sample of genotyped individuals, allowing us to implement a GWAS. The UKB also collected coordinate information on places of birth and current residence (at the time of the survey), which are required to construct a migration measure (described later). Of the respondents who completed the study ($n = 502{,}505$), we excluded those with no migration distance data ($n = 61{,}672$) and those of non-European ancestries ($n = 50{,}024$). After additional quality control, 359,571 samples remained.[6]

be combined into a polygenic index (PGI) at the individual (respondent) level. A few studies have compared educational attainment PGIs between migrants and nonmigrants (Abdellaoui et al. 2022; Abdellaoui et al. 2019; Belsky et al. 2019; Belsky et al. 2016), but we are unaware of research performing GWAS for migration outcomes.

[5] Migration and SES (and probably health outcomes) are distal phenotypes, suggesting that proximate variables mediate the associations of genetic traits with these outcomes. However, the presence of mediating factors does not eliminate the value of this study's unique contributions discussed in the following passage in this paragraph.

[6] For example, we randomly selected individuals among those in second-degree relative dyads by using KING (https://www.kingrelatedness.com/) to calculate the genetic relatedness of all UKB respondents, as is standard in genetic analysis.

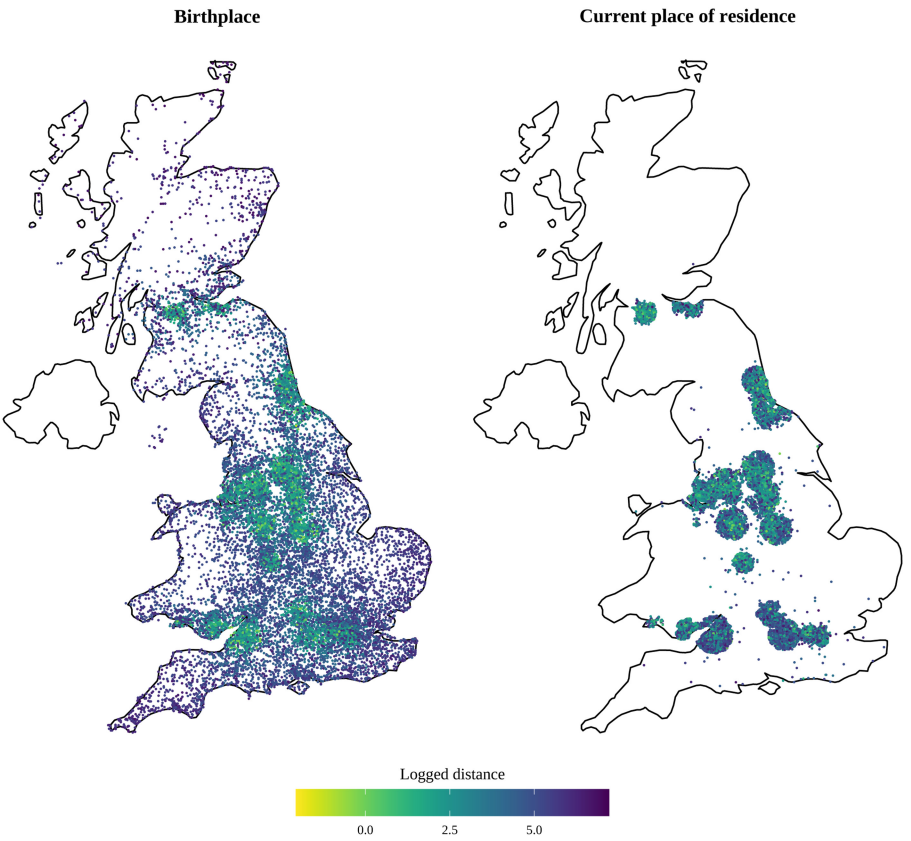

Fig. 1 Distribution of logged migration distance and geographic distribution of UKB participants. The analytic sample includes those of European ancestries with migration distance data. Logged distance indicates log-transformed travel distance (km) between birthplace and the current place of residence. The distribution of logged migration distance is available in Figure A1.

We used single-nucleotide polymorphisms (SNPs) as a genetic marker.[7] Our dependent variable is migration distance, representing the routing distance between the self-reported coordinates for respondent's places of birth and current residence. We measured migration distance as a continuous outcome because it does not require an arbitrary classification of respondents as (internal) migrants versus nonmigrants or long-distance versus short-distance migrants. Because preliminary analyses showed that more genetic variants correlated with logged migration distance than with migration distance, we focus on logged migration distance.[8] Additionally, we included the following control variables: age; sex; the type of chip used for genotyping; and the first 20 principal components, which account for population structure–related confounding (Price et al. 2006).

[7] SNP is a genetic variation in a single base pair at a specific location in DNA.

[8] Results of non-log-transformed migration distance are similar to our main findings (see section 2 of the online appendix).

We performed a GWAS for logged migration distance using Hail, a software tool for genetic analysis (https://hail.is/). GWAS runs millions of regressions to investigate how the variation of an outcome variable is associated with each genetic variant. Following conventions in the GWAS literature (e.g., Loh et al. 2015), we removed SNPs with a missing call rate greater than 0.01, a minor allele frequency less than 0.01, and a Hardy–Weinberg equilibrium test p value <1.0e–6. To control type I error, we used genomic control estimates (i.e., intercept) in linkage disequilibrium score (LDSC) regression (Bulik-Sullivan, Loh et al. 2015) to inflate standard errors for GWAS associations. Next, we calculated genetic correlations between our GWAS findings for logged migration distance and GWAS findings for 53 traits from other published studies. These traits include some genetic components that are direct (i.e., operate through inherited genetic variants) and some that are indirect (i.e., operate through the family environment) (Wu et al. 2021). This decomposition of genetic correlation allowed us to assess underlying mechanisms for the association between genetic variants and migration.[9] We used LDSC to estimate genetic correlations (Bulik-Sullivan, Finucane et al. 2015) and adjusted the significance cutoff using Bonferroni correction to account for multiple testing. GWAS summary statistics for the 53 traits are shown in Table A1 (tables and figures designated with an "A" are in the online appendix).

Results

Main Findings

Illustrating GWAS results, Figure 2 shows a Manhattan plot of 1,858 SNPs from 21 independent loci that reach the genome-wide significance level ($p < 5.0e–8$); genetic researchers use this very low p-value threshold to adjust for the hundreds of thousands of results estimated to control for false positive findings.[10] These SNPs are also associated with several SES and health outcomes. For example, outcomes associated with the SNP with the lowest p value in our migration analysis include educational attainment (Davies et al. 2016), cognitive performance (Lee et al. 2018), and anorexia nervosa (Peyrot and Price 2021). A measure of overall genetic contribution (SNP heritability) to logged migration distance is 0.0629 (standard error [SE] = 0.003): 6% of the variation is from commonly measured genetic variation.[11]

Figure 3 and Table A2 summarize genetic correlations (r_g) between logged migration distance and 53 traits. We find a strong positive genetic correlation between logged migration distance and educational attainment (r_g = 0.886); this level of genetic correlation is among the highest reported in the literature, exceeding that for

[9] One way to evaluate underlying mechanisms is to conduct a mediation analysis. However, in our case, a conventional mediation analysis is difficult to implement because we do not know the timing of migration. We therefore assess underlying mechanisms by decomposing genetic correlations into direct and indirect components.

[10] See Figure A2 for the quantile–quantile plot.

[11] To test potential mechanisms of the association between genetic variants and migration distance, we assessed whether sex and birth cohort (the 1940s, 1950s, and 1960s cohorts) moderate this relationship. We found no empirical evidence that these axes of social stratification moderate the relationship between genetic variants and migration distance (results available upon request).

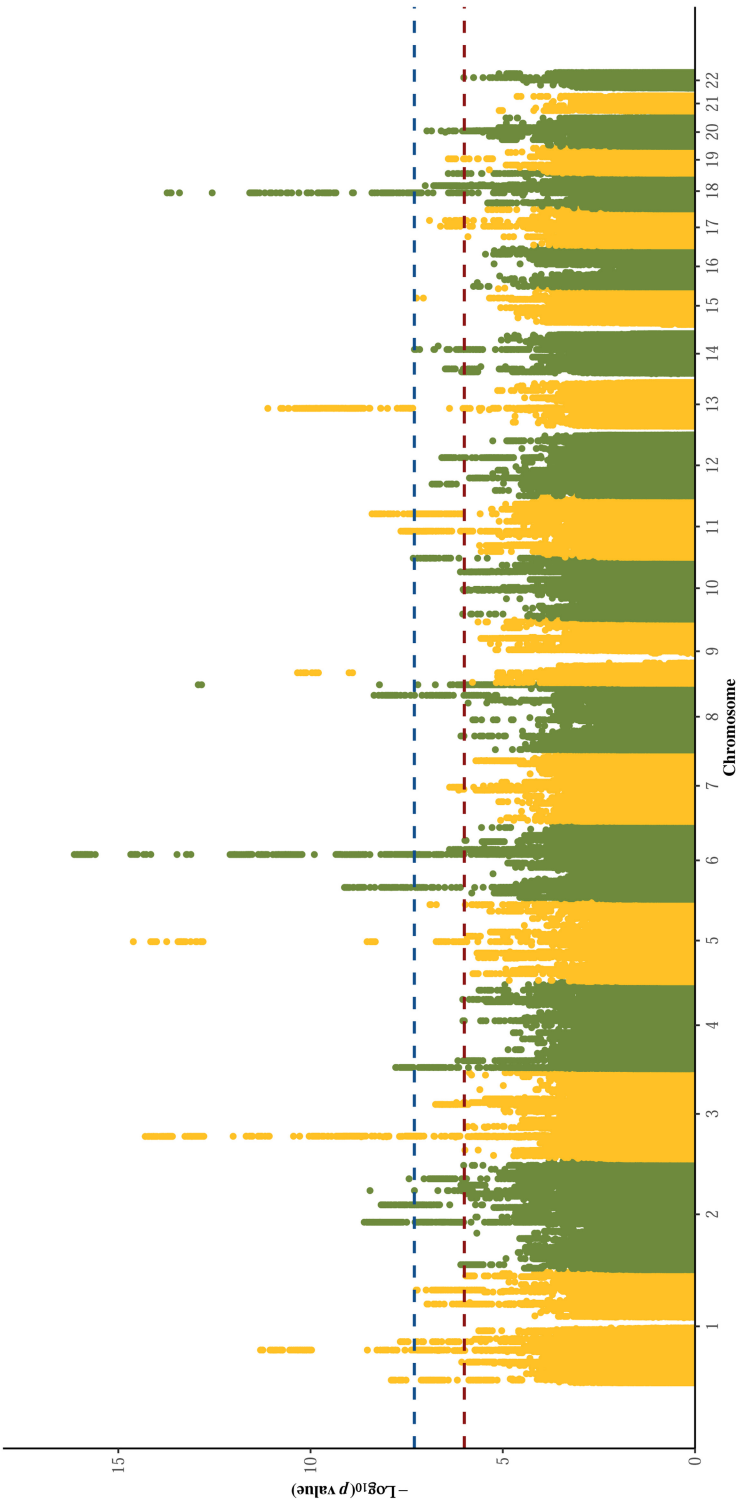

Fig. 2 Manhattan plot for logged migration distance. Dots represent the associations between SNP and migration distance. The dotted horizontal lines mark the genome-wide significance cutoff of 5.0e–8 (blue line) and a suggestive cutoff of 1.0e–6 (red line).

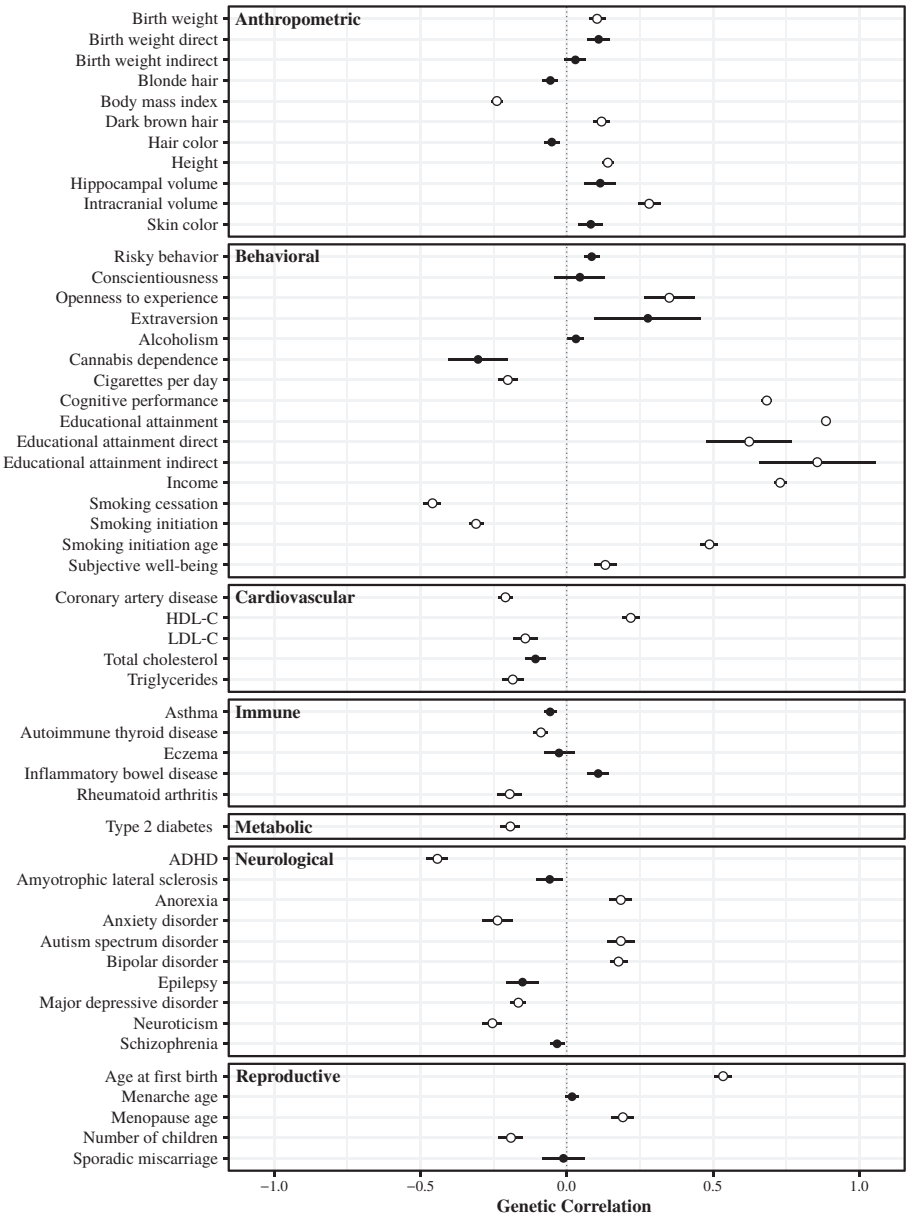

Fig. 3 Genetic correlations for logged migration distance. Circles and error bars indicate genetic correlation estimates and standard errors. Correlations significant at the 5% level after Bonferroni correction are highlighted as open circles. ADHD = attention-deficit/hyperactivity disorder. LDL-C and HDL-C = low- and high-density lipoprotein cholesterol, respectively.

cognitive performance. Further, our results of direct (i.e., own genetics) and indirect (i.e., parental/family genetics) components demonstrate strong genetic correlations with both dimensions but indicate a stronger genetic correlation with the indirect component ($r_g = 0.856$) than with the direct component ($r_g = 0.624$).

The results also reveal significant negative genetic correlations with several health-related issues, including coronary artery disease, Type 2 diabetes, major depressive disorder, neuroticism, and attention-deficit/hyperactivity disorder. Among fertility-related outcomes, genetic correlations with age at first birth and age at menopause are positive; the genetic correlation with the number of children is negative. Finally, several genetic correlations with health outcomes were unanticipated. Specifically, positive genetic correlations with anorexia nervosa, autism spectrum disorder, and bipolar disorder suggest higher genetic risks of these mental disorders among migrants relative to nonmigrants.[12]

Robustness Checks

Additional GWAS and Genetic Correlation Analyses

We conducted several robustness checks to assess the impacts of UKB's sampling design on our findings. Specifically, we investigated (1) the consequences of the overrepresentation of well-educated UK residents in the UKB (Munafò et al. 2018), (2) the impacts of the potential oversampling of health professionals,[13] and (3) the effects of sampling selection based on migration distance. To test the robustness of our findings on the first issue, we reimplemented GWAS while excluding those with professional education and college graduates. Our goal was to reduce the data's cases of migration for pursuing higher education. Similarly, we also ran GWAS excluding health professionals to eliminate the impacts of health professionals' migration into places around the medical assessment centers.[14] Regarding the third issue, we split the 22 catchment areas into two groups on the basis of place-specific median migration distance and performed GWAS separately for these two groups. A key sampling feature of the UKB is that only people living close to one of 22 assessment centers were asked to participate. This feature may truncate some internal migration distances in the full UK population. We split the data based on the place-specific migration distance distributions and examined the similarity of the results between the two subsamples. That is, we further truncated migration distance in each subsample and explored whether doing so would shape our results to gauge whether the

[12] Similarly, Figure A3 shows that genetic correlations between educational attainment and these health measures are also positive.

[13] We expect health professionals to be overrepresented in the sample because the UKB recruited individuals living close to one of 22 medical assessment centers.

[14] We identified health professionals using an employment history question in the UKB inquiring about paid jobs and apprenticeships held, in alignment with the international classification of health workers provided by the World Health Organization (https://www.who.int/publications/m/item/classifying-health-workers). We then excluded individuals categorized as health professionals and health associate professionals. Table A3 summarizes the job code in the UKB and the occupational classification.

unknown truncation due to the UKB sampling strategy is likely to have affected our main results.

Figures A4–A7 show that genetic correlations between migration distance and 53 traits are generally consistent across the ways we select the analytic sample. The correlation coefficient of genetic correlations between the sample with and without professional or college education is 0.98. Likewise, the correlation coefficient of genetic correlations between the sample with and without health professionals is 0.99. Further, the correlation coefficient of genetic correlations between our two subsamples based on migration distance is 0.97. Regression slopes in Figures A5–A7 are close to 1, ranging from 0.943 (SE = 0.022) to 1.165 (SE = 0.031). Hence, these results do not present empirical evidence that sampling selection issues in the UKB substantially alter our findings.

Within-Sibling Analyses

Differences in genetic ancestries have unignorable impacts on GWAS findings when genetic ancestries affect genetic variants and an outcome of interest. To account for this population stratification issue, we restricted the analytic sample to individuals of European ancestries and included genetic principal components in GWAS. However, principal components might not fully account for population stratification (Howe et al. 2022). To further eliminate the impacts of population stratification on our main findings, we conducted within-sibling GWAS, which compares genetic variants between siblings. This approach ensures that differences in genetic variants are not due to population stratification because siblings share genetic ancestries (Raffington et al. 2020).

Although within-sibling GWAS effectively reduces the threat of population stratification, this approach has limitations. First, within-sibling GWAS includes only UKB respondents whose siblings also participated in the UKB and therefore has a much smaller sample size (16,220 pairs and 32,440 individuals) than population GWAS. Second, within-sibling GWAS accounts for not only population stratification but also any other shared traits between siblings, such as the family of origin's socioeconomic background and childhood neighborhood environments. Because these shared traits explain some variance of an outcome measure, the remaining variance that genetic predispositions can explain is small in within-sibling GWAS. These limitations result in larger standard errors for genetic correlations in within-sibling GWAS than for population GWAS. Therefore, we primarily focused on sign concordance of genetic covariances for associations between genetic variants correlated with migration distance and other phenotypes. Finally, differences between population and within-sibling GWAS should be interpreted with caution. Consistent signs between population and within-sibling GWAS support our main findings with the least threat of population stratification. However, different signs do not imply that population stratification induces biased estimates in population GWAS because the within-sibling comparison accounts for all shared traits between siblings, including but not limited to population stratification.

Results of sign tests are presented in Table 1. Among the 33 traits that reached the 5% significance level after Bonferroni correction in the genetic correlation

Table 1 Estimated genetic covariances from population and within-sibling GWAS

Trait	Genetic Covariance		Same Sign?
	Population GWAS	Within-Sibling GWAS	
ADHD	–0.0580 (0.0052)***	–0.0154 (0.0128)	Yes
Age at First Birth	0.0352 (0.0022)***	0.0087 (0.0063)	Yes
Anorexia	0.0213 (0.0047)***	0.0121 (0.0127)	Yes
Anxiety Disorder	–0.0289 (0.0062)***	–0.0279 (0.0189)	Yes
Autism Spectrum Disorder	0.0223 (0.0059)***	–0.0146 (0.0200)	No
Autoimmune Thyroid Disease	–0.0116 (0.0033)***	–0.0099 (0.0135)	Yes
Bipolar Disorder	0.0288 (0.0050)***	0.0001 (0.0151)	Yes
Birth Weight	0.0091 (0.0025)***	0.0039 (0.0073)	Yes
Body Mass Index	–0.0300 (0.0025)***	0.0060 (0.0072)	No
Cigarettes per Day	–0.0153 (0.0021)***	–0.0015 (0.0063)	Yes
Cognitive Performance	0.0836 (0.0035)***	0.0188 (0.0081)*	Yes
Coronary Artery Disease	–0.0142 (0.0015)***	–0.0014 (0.0048)	Yes
Dark Brown Hair	0.0105 (0.0026)***	0.0015 (0.0072)	Yes
Educational Attainment	0.0887 (0.0030)***	0.0273 (0.0056)***	Yes
Educational Attainment, Direct	0.0428 (0.0063)***	0.0215 (0.0196)	Yes
Educational Attainment, Indirect	0.0498 (0.0062)***	0.0140 (0.0162)	Yes
HDL-C	0.0207 (0.0027)***	–0.0033 (0.0088)	No
Height	0.0279 (0.0041)***	–0.0116 (0.0112)	No
Income	0.0544 (0.0027)***	0.0138 (0.0064)*	Yes
Intracranial Volume	0.0401 (0.0060)***	0.0287 (0.0182)	Yes
LDL-C	–0.0122 (0.0032)***	–0.0136 (0.0108)	Yes
Major Depressive Disorder	–0.0111 (0.0018)***	–0.0132 (0.0054)*	Yes
Menopause Age	0.0202 (0.0040)***	0.0148 (0.0104)	Yes
Neuroticism	–0.0209 (0.0028)***	–0.0203 (0.0093)*	Yes
Number of Children	–0.0082 (0.0018)***	–0.0018 (0.0048)	Yes
Openness to Experience	0.0329 (0.0072)***	0.0228 (0.0218)	Yes
Rheumatoid Arthritis	–0.0248 (0.0037)***	–0.0218 (0.0125)†	Yes
Smoking Cessation	–0.0225 (0.0017)***	–0.0091 (0.0048)†	Yes
Smoking Initiation	–0.0222 (0.0019)***	–0.0064 (0.0044)	Yes
Smoking Initiation Age	0.0291 (0.0021)***	0.0041 (0.0058)	Yes
Subjective Well-being	0.0058 (0.0017)***	0.0020 (0.0057)	Yes
Triglycerides	–0.0203 (0.0035)***	–0.0032 (0.0112)	Yes
Type 2 Diabetes	–0.0148 (0.0026)***	0.0036 (0.0076)	No

Notes: Data are restricted to the 33 phenotypes showing significant genetic correlations with migration distance after Bonferroni correction in the main analysis. Standard errors are shown in parentheses. ADHD = attention-deficit/hyperactivity disorder. LDL-C and HDL-C = low- and high-density lipoprotein cholesterol, respectively.

†$p < .10$; *$p < .05$; ***$p < .001$

analysis, genetic covariances of 28 traits (85% of the 33 traits) from within-sibling GWAS have signs consistent with those from population GWAS ($p = 6.6 \cdot 10^{-5}$, binomial test).[15] These results suggest that significant genetic correlations for

[15] Because there are positive and negative signs, we expect that half of the genetic covariances from within-sibling GWAS show inconsistent signs with the genetic covariances from population GWAS if the signs of genetic covariances from within-sibling GWAS are random. Therefore, the null hypothesis in this

migration distance with SES and health in the main analyses are generally robust to population stratification.

By contrast, the signs of genetic covariances of the other five traits from within-sibling GWAS differ from those in population GWAS. These traits are autism spectrum disorder, body mass index, HDL cholesterol, height, and Type 2 diabetes. Although these differences are probably due to population stratification in population GWAS, we cannot reject an alternative scenario that other shared traits between siblings alter the signs of genetic covariances of these characteristics. Furthermore, these genetic covariances are not statistically significant even without Bonferroni correction, suggesting that the inconsistent signs may result from low statistical power and imprecise estimation. Because most of the traits show consistent directions of genetic covariances across within-sibling and population GWAS, we conclude that within-sibling GWAS does not present empirical evidence casting doubt on our main findings in the genetic correlation analysis.

Analyses With U.S. Data

To further validate our findings, we conducted similar analyses with different samples. However, we are unaware of datasets that provide a migration distance measure and genetic data with a sufficiently large sample size to implement a GWAS and genetic correlation analysis.[16] As an alternative, we used our GWAS findings to create a migration distance polygenic index (PGI): a summary measure representing cumulative correlations of independent genomic loci of small correlations with migration distance. We then assessed how migration distance PGI is associated with migration distance, health, SES, and skills in a U.S. population.[17] On the basis of our main findings, we expected that those with genetic variants correlated with longer migration distances (i.e., higher migration distance PGIs) move longer distances, are healthier, and have more socioeconomic resources than those with lower migration distance PGIs. Such results would provide additional support validating the findings in the GWAS and genetic correlation analysis.

We analyzed data from the National Longitudinal Study of Adolescent to Adult Health (Add Health) and the Health and Retirement Study (HRS), which collected genetic data. These datasets allowed us to construct migration distance PGIs. Further, because these datasets cover different age and birth cohorts, we could assess whether the associations of the migration distance PGI with health, SES, and skills depend on these demographic characteristics. We used Add Health Wave I to explore

binomial test is that the probability that the signs of genetic covariances from within-sibling GWAS are consistent with those from population GWAS is 0.5.

[16] For example, Add Health provides migration distance information and genetic data, but the sample size ($N = 4{,}508$ after quality checks) is too small to run a GWAS and genetic correlation analysis.

[17] To create the migration distance PGI, we used the results of the logged migration distance GWAS of UKB data. Following the standard procedure to construct PGI, we clumped SNPs using Phase 3 European samples from the 1,000 Genomes Project as linkage disequilibrium reference. The linkage disequilibrium window size and a pairwise R^2 threshold were set at 1 megabase (Mb) and 0.1, respectively. We did not use p value thresholding for variant selection. We calculated migration distance PGI with PRSice-2 software (Choi and O'Reilly 2019) and standardized with a mean of 0 and a variance of 1 in downstream analyses.

Table 2 Estimated relationships between logged migration distance PGI and geographic mobility distance in Add Health Waves I and III

	Distance Between Waves I and III	
Variable	Standardized	Log-Transformed
Migration PGI (standardized)	0.036*	0.148†
	(0.018)	(0.079)
Number of Observations	4,688	4,688

Notes: Logged migration distance PGI is standardized at a mean of 0 and standard deviation of 1. The sample is restricted to those of European ancestries. Heteroskedasticity-robust standard errors are shown in parentheses. Additional controls include the first 20 principal components, a dummy variable for age in Wave III, and a dummy variable for sex. Additional controls are not shown.

†$p < .10$; *$p < .05$

the association between the migration distance PGI and phenotypic traits among adolescents.[18] Additionally, we analyzed Add Health Wave IV and the 2012 round of HRS, which collected data for young adults and older people, because these survey waves collected completed genetic data.

Migration distance is measured by the distance of locations between Waves I and III in Add Health.[19] We created health, SES, and skill measures by using Add Health Waves I and IV and the 2012 round of HRS. Our health outcome measures include self-reported health, height, body mass index, and depression (assessed with the Center for Epidemiologic Studies Depression Scale). For Wave I respondents, we also used picture vocabulary test scores and grades in English, math, social studies, and science to measure respondents' abilities and skills. With Add Health Wave IV and HRS data, we measured respondents' SES as educational attainment and log-transformed individual and household income (Table A4 details the operationalizations of these outcome measures).

Table 2 summarizes the results of the association between migration distance PGI and the location distance between Add Health Waves I and III. Net of age, sex, and the first 20 principal components, a higher migration distance PGI is significantly associated with a longer migration distance. This finding suggests that genetic variants correlated with a longer migration distance among the UKB participants are also associated with a longer migration distance among the Add Health participants.

We then examined the associations with health, SES, and skills. Table 3 demonstrates that net of age, sex, and the first 20 principal components, a higher migration distance PGI is associated with better health and higher skills and SES, regardless of age groups and birth cohorts. These results are consistent with the positive genetic

[18] Add Health Wave II also provides data for adolescents, but the sample size is somewhat smaller in Wave II than Wave I.

[19] Add Health also provides the location distance between Waves I and II and between Waves II and III. However, only small variations in the location distance exist between Waves I and II because Wave II collected data one or two years after Wave I. Further, the location distances between Waves II and III are similar to those between Waves I and III, but Wave II has fewer observations than Wave I. Therefore, we used the location distance between Waves I and III.

Table 3 Estimated relationships between logged migration distance PGI and phenotypes in the U.S. data

	Add Health Wave I (adolescents)	Add Health Wave IV (adults)	HRS (older adults)
Variable	(1)	(2)	(3)
A. Health Outcomes			
Self-reported health	0.039**	0.062***	0.137***
	(0.013)	(0.013)	(0.020)
Height	0.012**	0.013***	0.004**
	(0.004)	(0.003)	(0.001)
Body mass index	–0.182**	–0.458***	–0.199†
	(0.062)	(0.105)	(0.111)
Depression (CES-D scale)	–0.381***	–0.196**	–0.144***
	(0.106)	(0.069)	(0.037)
B. SES and Skills			
Completed high school	—	0.016***	0.051***
	—	(0.003)	(0.006)
Completed four-year college	—	0.075***	0.105***
	—	(0.007)	(0.008)
Logged personal income	—	0.163***	0.154*
	—	(0.040)	(0.073)
Logged household income	—	0.055***	0.134***
	—	(0.012)	(0.019)
Picture vocabulary test score	1.969***	—	—
	(0.174)	—	—
Grade in English	0.090***	—	—
	(0.017)	—	—
Grade in math	0.100***	—	—
	(0.018)	—	—
Grade in social studies	0.123***	—	—
	(0.019)	—	—
Grade in science	0.127***	—	—
	(0.018)	—	—

Notes: Logged migration distance PGI is standardized at a mean of 0 and a standard deviation of 1. The sample is restricted to those of European ancestries. Heteroskedasticity-robust standard errors are shown in parentheses. Additional controls include the first 20 principal components, the age fixed effect, and a dummy variable for sex. Additional controls are not shown. CES-D = Center for Epidemiologic Studies Depression Scale.

†$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$

correlations between migration distance and health, SES, and skills in the main findings. Overall, the additional analyses with two U.S. datasets suggest that our UKB results generalize to other contexts.

Discussion

This study provides novel assessments of migration selection using genetic analytic tools. We found many genetic variants associated with logged migration distance. Because genetic variants are not affected by migration, these results provide direct

evidence of genetic migration selection: those with certain genetic variants are more likely to migrate than those without these variants. These findings support Palloni and Arias' (2004) speculation of the presence of migration selection at the genetic level.

Further, we found that genetic variants correlated with migration distance are also associated with many dimensions of SES and health outcomes. These results imply that migration selection at the genetic level is tied to many other characteristics. The positive genetic correlations with educational attainment, income, and cognitive performance suggest that skilled individuals are more likely to migrate and that long-distance migrants pursue better educational and occupational opportunities. These results show that the Roy model (Borjas 1987; Borjas et al. 1992) and the law of migration (Ravenstein 1885) have implications for the genetic profiles of migrants compared with nonmigrants. In the case of educational attainment–related genetics, our decomposition of genetic correlations into direct and indirect components provides some insights into mechanisms: genetics correlated with family environments related to higher educational attainment contribute to genetic migration selection more than genetics correlated with own skills and abilities for successful educational attainment.

One major advantage of genetic measures is the wide coverage of genetic correlations, especially with health outcomes. This feature allows us to examine a broader set of outcomes that are rarely available in most datasets. Indeed, this wide coverage in genetic correlations provides several important theoretical implications for the healthy migrant hypothesis (Jasso et al. 2004; Palloni and Arias 2004; Palloni and Morenoff 2006). First, the healthy migrant hypothesis is valid for health conditions and risks that people may not perceive before migration. For example, genetic correlations with chronic diseases usually appearing at middle or older ages (e.g., coronary artery disease) uncover the likelihood of migration selection in latent health risks, given that internal migration in the United Kingdom is concentrated at young adult ages (Bernard et al. 2016). By contrast, our findings also provide nuance to the healthy migrant hypothesis and suggest the need for additional research. Specifically, significant positive genetic correlations between migration distance and bipolar disorder and anorexia nervosa suggest that those with higher genetic risks of these mental conditions are more likely to migrate. These genetic correlations are counterintuitive to the theoretical explanation that healthy individuals (in this case, those with lower risks of mental disorders) are more likely to migrate. One possible interpretation for these unanticipated results is that those with a high genetic risk of these mental conditions may be skilled individuals. This scenario is consistent with our genetic correlations between educational attainment and these mental disorders, as well as prior epidemiological research (MacCabe et al. 2010; Tiihonen et al. 2005). Because skilled individuals are more likely to migrate, those with genetic variants correlated with a higher risk of these mental disorders may also be more likely to migrate. Overall, these findings lead us to hypothesize that the healthy migrant hypothesis may not apply to some mental conditions, which are positively correlated with SES.

We acknowledge several limitations in this study. First, UKB is not a nationally representative survey and does not provide sampling weights to adjust the unique sampling strategy. Although we conducted many robustness checks to assess the potential impacts of sampling selection, subsequent research with large, nationally representative data can further explore the consequences of this limitation.

Second, we excluded respondents of non-European ancestries to increase ancestral homogeneity. This exclusion limits the generalizability of our findings of genetic migration selection. Finally, the genetic correlation analysis does not fully reveal the underlying mechanisms of migration selection. The result of a higher genetic correlation with the indirect component of educational attainment than with the direct component provides insights into the mechanisms, but we remain uncertain about what specific family or nurturing environments contribute to migration selection by educational attainment.

Despite these limitations, our study makes valuable contributions to the study of migration selection, which is typically constrained by data availability. By leveraging the unique feature of genetic measurements, we documented the presence of migration selection at the genetic level. Although we showed that genetic migration selection is generally consistent with theories and empirical evidence in migration selection, we also found genetic migration selection counter to our theoretical expectation. These unanticipated results generate novel hypotheses, and subsequent tests of the hypothesis will shed light on understudied aspects of migration selection. ■

Acknowledgments The authors acknowledge the use of the facilities of the Center for Demography of Health and Aging (P30 AG016266) and the Center for Demography and Ecology (P2C HD067873) at the University of Wisconsin–Madison. We thank the University of Wisconsin's Social Genomics Research Group members for helpful comments. An earlier version of this paper was presented at the 2021 annual meeting of the Population Association of America and at the National Institute on Aging and the 2021 Integrating Genetics and Social Sciences Conference (R13-AG062366). This research was conducted using the UK Biobank Resource under application number 57284 (http://www.ukbiobank.ac.uk/). GWAS summary statistics for (logged) migration distance are available at http://qlu-lab.org/data.html. Q. Lu and J. M. Fletcher codirected this research.

References

Abdellaoui, A., Dolan, C. V., Verweij, K. J. H., & Nivard, M. G. (2022). Gene–environment correlations across geographic regions affect genome-wide association studies. *Nature Genetics, 54,* 1345–1354.

Abdellaoui, A., Hugh-Jones, D., Yengo, L., Kemper, K. E., Nivard, M. G., Veul, L., . . . Visscher, P. M. (2019). Genetic correlates of social stratification in Great Britain. *Nature Human Behaviour, 3,* 1332–1342.

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review, 102,* 1832–1856.

Akresh, I. R., & Frank, R. (2008). Health selection among new immigrants. *American Journal of Public Health, 98,* 2058–2064.

Altman, C. E., Van Hook, J., & Hillemeier, M. (2016). What does self-rated health mean? Changes and variations in the association of obesity with objective and subjective components of self-rated health. *Journal of Health and Social Behavior, 57,* 39–58.

Belsky, D. W., Caspi, A., Arseneault, L., Corcoran, D. L., Domingue, B. W., Harris, K. M., . . . Odgers, C. L. (2019). Genetics and the geography of health, behaviour and attainment. *Nature Human Behaviour, 3,* 576–586.

Belsky, D. W., Moffitt, T. E., Corcoran, D. L., Domingue, B., Harrington, H. L., Hogan, S., . . . Caspi, A. (2016). The genetics of success: How single-nucleotide polymorphisms associated with educational attainment relate to life-course development. *Psychological Science, 27,* 957–972.

Beltrán-Sánchez, H., Palloni, A., Riosmena, F., & Wong, R. (2016). SES gradients among Mexicans in the United States and in Mexico: A new twist to the Hispanic paradox? *Demography, 53,* 1555–1581.

Bernard, A., Bell, M., & Charles-Edwards, E. (2016). Internal migration age patterns and the transition to adulthood: Australia and Great Britain compared. *Journal of Population Research, 33,* 123–146.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *American Economic Review, 77,* 531–553.

Borjas, G. J., Bronars, S. G., & Trejo, S. J. (1992). Self-selection and internal migration in the United States. *Journal of Urban Economics, 32,* 159–185.

Bostean, G. (2013). Does selective migration explain the Hispanic paradox? A comparative analysis of Mexicans in the U.S. and Mexico. *Journal of Immigrant and Minority Health, 15,* 624–635.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., . . . Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics, 47,* 1236–1241.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, . . . Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics, 47,* 291–295.

Choi, S. W., & O'Reilly, P. F. (2019). PRSice-2: Polygenic risk score software for biobank-scale data. *GigaScience, 8,* giz082. https://doi.org/10.1093/gigascience/giz082

Crimmins, E., Kim, J. K., & Vasunilashorn, S. (2010). Biodemography: New approaches to understanding trends and differences in population health and mortality. *Demography, 47*(Suppl. 1), S41–S64.

Crimmins, E. M., Soldo, B. J., Kim, J. K., & Alley, D. E. (2005). Using anthropometric indicators for Mexicans in the United States and Mexico to understand the selection of migrants and the "Hispanic paradox." *Social Biology, 52,* 164–177.

Darlington, F., Norman, P., & Gould, M. (2015). Health and internal migration. In D. P. Smith, N. Finney, K. Halfacree, & N. Walford (Eds.), *Internal migration: Geographical perspective and processes* (pp. 113–128). Surrey, UK: Ashgate Publishing. Retrieved from http://eprints.whiterose.ac.uk/89994/

Davies, G., Marioni, R. E., Liewald, D. C., Hill, W. D., Hagenaars, S. P., Harris, S. E., . . . Deary, I. J. (2016). Genome-wide association study of cognitive functions and educational attainment in UK Biobank ($N = 112\ 151$). *Molecular Psychiatry, 21,* 758–767.

Feliciano, C. (2005). Educational selectivity in U.S. immigration: How do immigrants compare to those left behind? *Demography, 42,* 131–152.

Feliciano, C. (2020). Immigrant selectivity effects on health, labor market, and educational outcomes. *Annual Review of Sociology, 46,* 315–334.

Fuller-Thomson, E., Brennenstuhl, S., Cooper, R., & Kuh, D. (2015). An investigation of the healthy migrant hypothesis: Pre-emigration characteristics of those in the British 1946 birth cohort study. *Canadian Journal of Public Health / Revue Canadienne de Santé Publique, 106,* e502–e508. https://doi.org/10.17269/cjph.106.5218

Grol-Prokopczyk, H., Freese, J., & Hauser, R. M. (2011). Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior, 52,* 246–261.

Harris, K. M., & Schorpp, K. M. (2018). Integrating biomarkers in social stratification and health research. *Annual Review of Sociology, 44,* 361–386.

Howe, L. J., Nivard, M. G., Morris, T. T., Hansen, A. F., Rasheed, H., Cho, Y., . . . Davies, N. M. (2022). Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature Genetics, 54,* 581–592.

Jasso, G., Massey, D. S., Rosenzweig, M. R., & Smith, J. P. (2004). Immigrant health: Selectivity and acculturation. In N. B. Anderson, R. A. Bulatao, & B. Cohen (Eds.), *Critical perspectives on racial and ethnic differences in health and late life* (pp. 227–266). Washington, DC: National Academies Press.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., . . . Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics, 50,* 1112–1121.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., . . . Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics, 47,* 284–290.

Lu, Y. (2008). Test of the "healthy migrant hypothesis": A longitudinal analysis of health selectivity of internal migration in Indonesia. *Social Science & Medicine, 67,* 1331–1339.

Lu, Y. (2010). Rural-urban migration and health: Evidence from longitudinal data in Indonesia. *Social Science & Medicine, 70,* 412–419.

MacCabe, J. H., Lambe, M. P., Cnattingius, S., Sham, P. C., David, A. S., Reichenberg, A., . . . Hultman, C. M. (2010). Excellent school performance at age 16 and risk of adult bipolar disorder: National cohort study. *British Journal of Psychiatry, 196,* 109–115.

Mehta, N. K., & Elo, I. T. (2012). Migrant selection and the health of U.S. immigrants from the former Soviet Union. *Demography, 49,* 425–447.
Morey, B. N., Bacong, A. M., Hing, A. K., de Castro, A. B., & Gee, G. C. (2020). Heterogeneity in migrant health selection: The role of immigrant visas. *Journal of Health and Social Behavior, 61,* 359–376.
Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Davey Smith, G. (2018). Collider scope: When selection bias can substantially influence observed associations. *International Journal of Epidemiology, 47,* 226–235.
Nauman, E., VanLandingham, M., Anglewicz, P., Patthavanit, U., & Punpuing, S. (2015). Rural-to-urban migration and changes in health among young adults in Thailand. *Demography, 52,* 233–257.
Palloni, A., & Arias, E. (2004). Paradox lost: Explaining the Hispanic adult mortality advantage. *Demography, 41,* 385–415.
Palloni, A., & Morenoff, J. D. (2006). Interpreting the paradoxical in the Hispanic paradox. *Annals of the New York Academy of Sciences, 954,* 140–174.
Peyrot, W. J., & Price, A. L. (2021). Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS. *Nature Genetics, 53,* 445–454.
Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38,* 904–909.
Raffington, L., Mallard, T., & Harden, K. P. (2020). Polygenic scores in developmental psychology: Invite genetics in, leave biodeterminism behind. *Annual Review of Developmental Psychology, 2,* 389–411.
Rauscher, E., & Oh, B. (2021). Going places: Effects of early U.S. compulsory schooling laws on internal migration. *Population Research and Policy Review, 40,* 255–283.
Ravenstein, E. G. (1885). The laws of migration. *Journal of the Statistical Society of London, 48,* 167–235.
Riosmena, F., Wong, R., & Palloni, A. (2013). Migration selection, protection, and acculturation in health: A binational perspective on older adults. *Demography, 50,* 1039–1064.
Ro, A., Fleischer, N. L., & Blebu, B. (2016). An examination of health selection among U.S. immigrants using multi-national data. *Social Science & Medicine, 158,* 114–121.
Rossouw, L., Bago d'Uva, T., & van Doorslaer, E. (2018). Poor health reporting? Using anchoring vignettes to uncover health disparities by wealth and race. *Demography, 55,* 1935–1956.
Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers, 3,* 135–146.
Rubalcava, L. N., Teruel, G. M., Thomas, D., & Goldman, N. (2008). The healthy migrant effect: New findings from the Mexican Family Life Survey. *American Journal of Public Health, 98,* 78–84.
Schmidt, R., Kristen, C., & Mühlau, P. (2022). Educational selectivity and immigrants' labour market performance in Europe. *European Sociological Review, 38,* 252–268.
Tiihonen, J., Haukka, J., Henriksson, M., Cannon, M., Kieseppä, T., Laaksonen, I., . . . Lönnqvist, J. (2005). Premorbid intellectual functioning in bipolar disorder and schizophrenia: Results from a cohort study of male conscripts. *American Journal of Psychiatry, 162,* 1904–1910.
Tolnay, S. E. (1998). Education selection in the migration of southern Blacks, 1880–1990. *Social Forces, 77,* 487–514.
Wilding, S., Martin, D., & Moon, G. (2016). The impact of limiting long term illness on internal migration in England and Wales: New evidence from census microdata. *Social Science & Medicine, 167,* 107–115.
Wu, Y., Zhong, X., Lin, Y., Zhao, Z., Chen, J., Zheng, B., . . . Lu, Q. (2021). Estimating genetic nurture with summary statistics of multigenerational genome-wide association studies. *Proceedings of the National Academy of Sciences, 118,* e2023184118. https://doi.org/10.1073/pnas.2023184118

Shiro Furuya (corresponding author)
furuya2@wisc.edu

Furuya • Department of Sociology, Center for Demography of Health and Aging, and Center for Demography and Ecology, University of Wisconsin–Madison, Madison, WI, USA; https://orcid.org/0000-0001-5683-8759

Liu • Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; https://orcid.org/0009-0001-9208-7902

Sun • Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI, USA; https://orcid.org/0009-0004-5682-2078

Lu • Center for Demography of Health and Aging, Department of Biostatistics and Medical Informatics, and Department of Statistics, University of Wisconsin–Madison, Madison, WI, USA; https://orcid.org/0000-0002-4514-0969

Fletcher • Center for Demography of Health and Aging, Center for Demography and Ecology, La Follette School of Public Affairs, Department of Population Health Science, and Department of Agricultural and Applied Economics, University of Wisconsin–Madison, Madison, WI, USA; https://orcid.org/0000-0001-8843-0563