

No. 2509

DNA Reveals the Growing Ancestral Diversity of the United States

Sam Trejo, Marissa Thompson

October 2025

DNA Reveals the Growing Ancestral Diversity of the United States

Sam Trejo^{1,2†}
Marissa Thompson³

[†]Send correspondence to samtrejo@princeton.edu

1. Department of Sociology, Princeton University; Princeton, NJ.
2. Office of Population Research, Princeton University; Princeton, NJ.
3. Department of Sociology, Columbia University; New York, NY.

Abstract:

Demographic research often seeks to assess changes to the characteristics of a population over time. However, many sociodemographic variables, like race/ethnicity, are self-reported measures that correspond to a subjective sense of identity; this makes it challenging to distinguish [i] changes to fixed characteristics of a population from [ii] changing social norms and patterns of self-identification. To address this issue, we utilize *genetic similarity proportions* (GSPs), which quantify the fraction of a person's DNA linked to various present-day reference populations. We analyze the dynamic relationship between race/ethnicity and GSPs across three American birth cohorts: 1945, 1980, and 2015. Our results highlight the growing ancestral diversity of the U.S. over time, including both within-race/ethnicity shifts in average GSPs and changes to the association between GSPs and racial/ethnic identification.

Introduction

In the two decades since the completion of the Human Genome Project (1), our ability to measure and understand human genetic variation has progressed remarkably. These genomic advances have yielded a plethora of scientific discoveries, from mapping the genetic architecture of a wide range of complex traits and diseases (2, 3) to understanding humanity's ancestral and evolutionary origins (4–6). However, while much attention has been paid to genomic discovery and historical human migration and admixture, relatively few studies leverage molecular genetic data to understand present-day demographic trends.

The racial/ethnic diversity of the United States has undoubtedly undergone substantial changes over the past century. In 1967, the year that the U.S. Supreme Court overturned state anti-miscegenation laws in *Loving v. Virginia*, interracial marriages comprised just 3% of all unions; today, they comprise approximately 17% of unions (and an even larger fraction of new births; (7)). Moreover, processes of fertility, mortality, and migration have produced changes in the population share of different racial/ethnic groups (8–10). However, racial/ethnic identity is a self-reported measure which corresponds to a subjective sense of group membership; that is, both race and ethnicity are social and political constructs (11, 12). Just as the underlying population of the U.S. has changed over time, so too have the social norms that govern racial/ethnic self-identification; notably, these changes are often gradual and typically lack a clear reference point. As a result, it has been virtually impossible – until now – to use traditional census and/or survey data to distinguish changes to individual characteristics, like a person's genealogical or genetic ancestry (13), from changing social norms and patterns of racial/ethnic self-identification.

Drawing on genotyped respondents from two nationally representative surveys of Americans born roughly a generation apart (14, 15), we construct *genetic similarity proportions* (16) – DNA-based measures not unlike those provided by popular at-home genetic ancestry tests – linked to four present-day reference populations (17, 18): Sub-Saharan Africa (Yoruba in Nigeria; P^{AFR}), Europe (Northern and Western Europeans in Utah; P^{EUR}), East Asia (Han Chinese in Beijing; P^{EAS}), and Indigenous America (Pima/Maya in Mexico P^{IAM}).¹ Genetic similarity proportions (GSPs; (16)), which are sometimes also referred to as *genetic ancestry proportions*, have a wide range of applications in human genomics (20–23) and are beginning to be utilized in the social sciences (24–26). Unlike self-identified racial/ethnic identity, which may change across contexts and over time (27, 28), an individual's genome – and, so too, any prespecified transformation of the genome – is stable across the life-course and directly transmitted from parents to their biological children. Thus, GSPs provide a reference point which allows us to compare individuals of similar ancestries (and, likely, physical appearance) who lived in different time periods and who were, therefore, subject to different social norms regarding racial/ethnic self-identification. Moreover, GSPs vary continuously, display ample variation even among members of the same self-identified racial/ethnic group, and have been shown to be associated with experiences of racialization by external observers (25, 26).

Note, the estimation of GSPs requires various decisions for which there are no right answers, and therefore any given set of GSP estimates do not represent an 'objective' measure of a person's ancestry – nonetheless, such measures, when constructed thoughtfully, represent a latent variable with useful properties for social and demographic inquiry. The specific GSPs

¹ In line with recent recommendations from the National Academies of Sciences, Engineering, and Medicine, we refer to the quantitative estimates of the genetic resemblance between individuals as *genetic similarity* (rather than *genetic ancestry*) (16). Some have argued that using the genetic ancestry to refer to supervised estimates from software like ADMIXTURE is a misnomer, as in most cases such estimates often actually describe a person's genetic similarity to various present-day (rather than ancestral) populations (19).

utilized in this study represent the top four axes of genetic variation among representative samples of Americans (see *Materials and Methods* for more details).

In this study, we empirically explore the dynamic relationship between racial/ethnic identity and GSPs over time. Our results reveal the complex and growing ancestral diversity of the United States. We observe within-race shifts in average GSPs which produce *ancestral diffusion*, as we term it, a process wherein the differences in GSPs between self-identified racial/ethnic groups slowly attenuate across successive generations. In addition, we observe temporal changes to the association between a person's GSPs and their expected racial/ethnic identity, allowing us to observe the social process of racial construction across generations.

Our study makes multiple notable contributions. First, we document previously hidden demographic changes to the U.S. population over time, which have important consequences for the accuracy of recent biomedical advances. Second, we highlight the limitations of efforts to measure intergenerational demographic change using only self-reported survey measures. Third, we highlight the complex and dynamic relationship between race/ethnicity and ancestry, which serves to undermine biological essentialism by illustrating how notions of race are constructed and reconstructed over time. Finally, our work contributes to a growing body of research that illustrates how GSPs can be used as a new methodological tool in the social and demographic sciences.

Results

We begin with a simple descriptive analysis examining changes over time to the distribution of GSPs among members of a given racial/ethnic group. In particular, we focus on the 1945 and 1980 U.S. birth cohorts; for the 1945 birth cohort ($N=9,636$), we utilize a survey-weighted subsample from the Health and Retirement Study (HRS; (14)). For the 1980 birth cohort ($N=7,689$), we utilize survey-weighted data from the National Longitudinal Study of Adolescent to Adult Health (Add Health; (15)). In addition, we account for selection into genotyping using inverse probability weighting. We have aligned our HRS and Add Health analytic samples and defined variables of interest so as to maximize comparability across the two datasets – see *Materials and Methods* for more details.

Figure 1 contains six panels, each of which displays a histogram of the 1945 distribution of a given GSP overlaid on a histogram of the 1980 distribution. Specifically, Panels A, B, and C display the distributions of Sub-Saharan African genetic similarity (P^{AFR}) among self-identified Black Americans, European genetic similarity (P^{EUR}) among self-identified White Americans, and Indigenous American genetic similarity (P^{IAM}) among self-identified Hispanic Americans; on the other hand, Panels D, E, and F display the same three distributions for non-Black, non-White, and non-Hispanic Americans, respectively. In Figure 1, we display the particular GSP that tends to best distinguish a given racial/ethnic group from the other racial/ethnic groups.² For instance, in the HRS, Black Americans have high average levels of Sub-Saharan African genetic similarity ($\bar{P}_{45}^{AFR}=0.80$), whereas non-Black Americans tend to have very low levels ($\bar{P}_{45}^{AFR}=0.01$). Similarly, Hispanic Americans have moderate levels of Indigenous American genetic similarity ($\bar{P}_{45}^{IAM}=0.41$), while non-Hispanic Americans have values near zero ($\bar{P}_{45}^{IAM}=0.01$). Finally, while

² Here, 'racial/ethnic group' describes the intersection of traditional racial categories (e.g., Black, White, Native American, Asian/Pacific Islander, etc.) and Hispanicity. Note, these racial/ethnic groups are constructed to be mutually exclusive (see *Materials and Methods*). For the remainder of the paper, when we use the terms 'Black Americans' and 'White Americans', we are describing the group of individuals who self-identify as Black or White, respectively, and who do *not* also self-identify as Hispanic. Finally, we use 'racial identification' and 'racial self-identification' interchangeably.

White Americans have levels of European genetic similarity close to one ($\bar{P}_{45}^{EUR}=0.98$), non-White Americans tend to have far lower – although still meaningful – levels ($\bar{P}_{45}^{EUR}=0.29$). Table S1 displays the average values of our four GSPs among each self-identified racial/ethnic group in both the HRS and Add Health.

Figure 1, taken altogether, reveals a fascinating pattern. In Panels A, B, and C, the 1980 Add Health distribution (displayed in blue) is shifted slightly to the left of the 1945 HRS distribution (displayed in red). That is, Black, White, and Hispanic Americans with comparatively lower levels of Sub-Saharan African, European, and Indigenous American genetic similarity, respectively, comprise a larger share of the 1980 birth cohort (compared to the 1945 birth cohort). This leftward shift is most pronounced for Indigenous American genetic similarity among Hispanic Americans, which decreased from an average of 0.41 in 1945 to 0.29 in 1980 (diff=0.128, $p<0.001$); notably, similar patterns are observed even among Hispanics who identify as Mexican-American, implying that these results are not simply driven by underlying demographic shifts in the ethno-national composition of the Hispanic population (see figures S2 and S3). Turning to panels C, D, and F of Figure 1, we see a similar pattern in reverse; that is, the 1980 Add Health distribution is shifted to the right of the 1945 HRS distribution. In particular, non-Black, non-White, and non-Hispanic Americans with comparatively higher levels of Sub-Saharan African, European, and Indigenous American genetic similarity, respectively, comprise a larger share of the 1980 birth cohort; this rightward shift is most pronounced for European genetic similarity among non-White Americans, which increased from an average of 0.29 in 1945 to 0.34 in 1980 (diff=0.053, $p<0.001$).

What factors could be driving the observed changes to the within-group distributions of GSPs over time? On one hand, such shifts may simply result from changes to the broader distribution of GSPs in the United States; for instance, as the number of generations from initial admixture events between European populations and both Sub-Saharan African and Indigenous populations grows, and as interracial marriage and reproduction becomes more and more common, it may simply be that there are a growing number of Americans born with GSP values that were uncommon in previous birth cohorts. At the same time, however, it is also possible that a portion of the observed within-group changes of GSPs over time are due to changing social patterns of racial/ethnic self-identification. To test this possibility, Figure 2 explores changes to the relationship between GSPs and racial/ethnic self-identification over time.

Panels A and B of Figure 2 each display a binned scatter plot with a local polynomial fit line. In Panel A, average Sub-Saharan African genetic similarity is plotted against the fraction of respondents who identify as Black; similarly, in Panel B, average Indigenous American genetic similarity is plotted against the fraction of individuals who identify as Hispanic. Each panel of Figure 2 contain individuals from the 1945 birth cohort (displayed in red) and the 1980 birth cohort (displayed in blue). In both panels, we observe a strong monotonic relationship between the focal GSP and racial/ethnic identity in 1945 and 1980; that is, individuals with greater Sub-Saharan African and Indigenous American genetic similarity are more likely to identify as Black and Hispanic, respectively – a reflection of the fact that a person’s GSPs are correlated with a range of social and physical characteristics relevant to racial/ethnic identity formation.

Among Black Americans, the relationship between GSP and identification has remained stable across the two birth cohorts (and, indeed, the 95% confidence intervals in Panel A overlap). In 1945 and in 1980, a person with $P^{AFR} \approx 0.4$ is predicted to be equally likely to identify as Black and as another race. However, among Hispanic Americans, the relationship has shifted to the left over time; HRS respondents with relatively low levels of Indigenous American genetic similarity (e.g., $0 < P^{IAM} < 0.3$) are far more likely to identify as Hispanic in 1980 than

similar respondents in Add Health. A person with $P^{IAM} \approx 0.2$ from the 1945 birth cohort is predicted to identify as Hispanic only about 60% of the time, whereas a person from the 1980 birth cohort with the same P^{IAM} value is predicted to identify as Hispanic nearly 90% of the time. This result suggests a shifting social boundary regarding Hispanic identification.

We also estimate changes to the average GSPs of the United States population overall, focusing on the 1945, 1980, and 2015 birth cohorts. As before, we use the HRS and Add Health, to represent the 1945 and 1980 birth cohorts, respectively. For the 2015 birth cohort, we leverage a demographic projection model that requires minimal assumptions and utilizes data from both Add Health and the American Community Survey (29); see *Materials and Methods* for more details. Table 1 (Columns 1-3) displays estimates of the population averages of our four GSPs, as well as estimates of the fraction of the population who identify as a given racial/ethnic group.³

From 1945 to 2015, American birth cohorts have experienced meaningful demographic changes, both in terms of average GSPs and racial/ethnic composition. Overall, the U.S. has become increasingly genetically diverse over time, as European genetic similarity (P^{EUR}) has decreased and Sub-Saharan African (P^{AFR}), Indigenous American (P^{IAM}), and East Asian (P^{EAS}) genetic similarity have all increased. Similarly, the 1945 U.S. birth cohort is overwhelmingly White (83%) and has high levels of European genetic similarity ($\bar{P}_{45}^{EUR}=0.86$); on the other hand, the 2015 birth cohort is just 53% White, with lower levels of European genetic similarity ($\bar{P}_{15}^{EUR}=0.72$).⁴ Over the same period, the fraction of United States births that were Black, Hispanic, and Other Races grew substantially, from 11%, 5%, and 2% to 15%, 24%, and 8%, respectively.

The concurrent trends in GSPs and racial/ethnic identity raises a key question: are the observed changes in average GSPs explained by the U.S.'s changing racial/ethnic composition, or do GSPs reveal novel demographic shifts? Table 1 (Columns 4-9) also presents results from a *Kitagawa decomposition* (30), which separates changes to a given GSP over time into the sum of two distinct components: [i] the expected change to the GSP given shifts in racial/ethnic demographics (holding average between-group differences in the GSP fixed), and [ii] a residual term which corresponds to changes in the average GSP within racial/ethnic groups. Notably, while the changes to African and European genetic similarity are – by and large – what would have been expected given the observed racial/ethnic changes, changes to Indigenous American genetic similarity are far smaller than expected. The changing racial/ethnic composition of the U.S. between 1945 and 2015 – especially the large increase in the Hispanic population share – would have been expected to produce a 350% increase in Indigenous American genetic similarity from its 1945 average ($\bar{P}_{45}^{IAM}=0.25$); however, substantial changes to the distributions of GSPs within racial/ethnic groups partially offset this expected growth (see, for instance, Panel C of Figure 1), and Indigenous American genetic similarity of the 2015 birth cohort ($\bar{P}_{15}^{IAM}=0.75$) increased just 200% from its 1945 average (diff=0.05, $p<0.001$). Finally, to more explicitly quantify growing individual-level genetic diversity, we estimate average changes to a person's single highest GSP (as opposed to their other three GSPs) – see Table S2. While in the 1945 birth cohort, the average value of a person's single highest GSP was 0.95, it had decreased to 0.92 by the 1980 birth cohort (diff=0.029, $p<0.001$).

Discussion

³ In addition, Figure S1 uses GSP deciles to display distributional (rather than simply average) changes in GSPs from 1945 to 1980.

⁴ Importantly, a birth cohort's racial/ethnic demographics is defined retrospectively (once its members reach adulthood and have acquired and solidified racial/ethnic identities).

Our empirical results highlight the complex and dynamic relationship between ancestry and racial/ethnic identity over time. From 1945 to 1980 (and even to 2015), U.S. birth cohorts have become increasingly genomically diverse, as measured by GSPs – even within self-identified racial/ethnic groups. While a portion of these changes over time are explained by shifts in the overall racial/ethnic composition of the population, within-group changes in average GSPs have also played an important role. In particular, we find evidence for *ancestral diffusion*: a process wherein the genomic differences between self-identified racial/ethnic groups slowly attenuate across successive generations. Over time, Black Americans and Hispanic Americans have grown to have lower levels of Sub-Saharan African and Indigenous American genetic similarity, respectively; at the same time, both groups have grown to have increased levels of European genetic similarity. There is also notable variation in how fast these changes are occurring for different racial/ethnic groups; Hispanic Americans, for instance, are experiencing the most rapid and substantial within-group GSP changes, driven (in part) by a changing relationship between Indigenous American genetic similarity and Hispanic self-identification. These diffusion processes will likely continue, though only time will tell precisely what the coming generations will bring.

In addition, our results regarding the shifting relationship between race/ethnicity and ancestry inform an expansive literature on the social construction of race. Historically, conceptualizations of racial/ethnic categories were biological in nature. However, today such categories are widely understood to be socially constructed using a range of physical and social characteristics (28). Nonetheless, genetically essentialist views – the belief that DNA differences divide humans into discrete and biologically distinct racial groups – still pervade both the academy and broader society (31–34). At first blush, the sizable correlation that exists between a person’s racial/ethnic identity and their GSPs may seem like a challenge to the social constructivist understanding of race. However, processes of social (re)construction often involve humans layering meaning onto the physical, natural, and biological features of our world – for instance, dividing the surface of the Earth into distinct geopolitical entities (i.e., countries). In the case of racial/ethnic categories, it is the expansive family tree of humanity that has become imbued with social meaning and stigma (33, 34). Crucially, our results empirically highlight how the relationship between racial/ethnic identity and ancestry is inherently variable – as racial/ethnic schemas grow and evolve over time, the correspondence between race/ethnicity and ancestry also changes. Thus, our analysis helps to empirically and conceptually distinguish race/ethnicity, the social process of racial/ethnic categorization and identification, from ancestry, the pre-social family tree shared by all of humanity. Our results demonstrate that the racial/ethnic composition of future populations will not be driven solely by patterns of assortative mating and fertility that shape genealogical and genetic ancestral descent, but also by broader social changes regarding the rules and tendencies that govern self-identification.

Moreover, our findings highlight how social institutions and norms (such as the legalization and normalization of interracial marriage) can leave an enduring imprint on the demographic characteristics of a population. This can be seen with, for example, changes in the distribution of Sub-Saharan African genetic similarity among self-identified Black Americans. In the 1945 birth cohort, there are very few Black Americans with $0.4 < P^{AFR} < 0.6$; however, by the 1980 cohort, this population has grown substantially. It is beyond the scope of this study to definitively state how this change occurred, but a rise in childbirths from interracial marriages between Black and non-Black Americans following *Loving v. Virginia* may be an important part of the story.

Our results also have important implications for the biomedical sciences. The allele frequencies of genetic variants that impact various traits, like cystic fibrosis (35) and sickle cell disease (36), vary on average as a function of ancestry; thus, as the ancestral characteristics of the U.S. population change, so too may the prevalence and risk for certain diseases. In addition, the U.S.'s growing levels of Sub-Saharan African, Indigenous American, and East Asian genetic similarity have important implications for the accuracy of rapidly advancing genomic tools for complex traits, such as polygenic scores. The highly skewed ancestral composition of genome-wide association studies (37) has created large gaps in the accuracy of genetic risk prediction between individuals of European ancestries and individuals of non-European ancestries (20, 38). Our results suggest that the average accuracy of current polygenic scores may decline over time, as the overall target population becomes less genetically similar to the discovery samples of existing genomic studies. Moreover, in order to effectively design studies to reduce this so-called portability problem, researchers must have a representative understanding of the genetic diversity of their population of interest – as well as how that genetic diversity is changing. Benchmarking these changes using conventional genomic data sources is challenging due to the fact that most large-scale biobanks lack a sampling frame and therefore likely suffer from selection and ascertainment biases (39, 40).

Finally, by highlighting the limitations of self-reported racial/ethnic identity measures, our work contributes to a growing body of research that argues GSPs represent a useful methodological tool in the social sciences; importantly, GSPs represent a complement to, and not a replacement of, existing survey strategies for measuring racial-ethnic change. Understanding changing patterns of racial/ethnic identification – for instance, why certain Latin American populations are increasingly self-identifying as Indigenous (9) – is an interesting and important goal in and of itself. But our ability to make scientific progress in our understanding of racial/ethnic mobility and inequality is only as good as our capacity to accurately measure a population over time; however, a person's racial/ethnic identity – unlike their GSPs – may vary over time and is imperfectly transmitted from parents to children. This makes it difficult (or even impossible) to accurately observe key demographic processes – like, for instance, the intergenerational economic and educational mobility of Hispanic immigrants in the U.S. (27) – by tracking racial/ethnic groups through repeated cross-sections of the population (such as the U.S. census). In contrast, GSPs, may serve as a fixed reference point from which to view social and demographic change.

Materials and Methods

Datasets

We utilize data from two nationally representative longitudinal studies: The Health and Retirement Study (HRS; $N=10,819$) and The National Longitudinal Study of Adolescent to Adult Health (Add Health; $N=8,162$). The HRS is a nationally representative survey of Americans above the age of 50, which began in 1992 and has been repeated biennially since. We draw upon respondents from the following four of the study’s cohorts: HRS (born 1931-1941), War Babies (1942-1947), Early Baby Boomers (1948-1953), and Mid Baby Boomers (born 1954-1959). These HRS cohorts are survey weighted to be representative of the U.S. population born 1931-1959, with 1945.12 as the average birth year. Beginning in 2006, consenting HRS respondents provided saliva samples and were genotyped using the Illumina Human Omni-2.5 Quad BeadChip. To account for the HRS’s repeated longitudinal sampling design, we utilize person-level sampling weights from the wave that each HRS respondent was first empaneled.

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is a nationally representative survey of students who were in 7th through 12th grades during the 1994-1995 school year. The birth years of Add Health respondents range from 1974 to 1983, with 1979.17 as the survey weighted average birth year. In 2008 (Wave IV), consenting Add Health respondents provided saliva samples and were genotyped using one of two Illumina platforms – the Illumina Human Omni-1-Quad BeadChip and the Illumina Human Omni-2.5 Quad BeadChip. To account for Add Health’s complex survey design, we utilize the Wave IV grand sampling weights.

While Add Health respondents were empaneled at an average age of about 16, HRS respondents were empaneled between the ages of 51 and 61. For this reason, in order to ensure comparability of the underlying populations sampled by the HRS and Add Health, we restrict our analytic samples to U.S.-born individuals. In both studies, over 80% of eligible respondents consented to genotyping and rigorous quality control procedures were applied to the genotype data. Table S3 provides detailed descriptive statistics for our combined analytic sample.

Genetic Similarity Proportions

We use supervised ADMIXTURE (41–43), a maximum likelihood estimation approach, to construct genetic similarity proportions (GSPs) for each genotyped HRS and Add Health respondent. We model four ancestral populations using reference panels comprised of the following unrelated individuals from HapMap 3 (17) and the Human Genome Diversity Project (18): 83 Yoruba (Nigeria; P^{AFR}), 36 Northern/Western Europeans (Utah; P^{EUR}), 137 Han Chinese (Beijing; P^{EAS}), and 34 Pima/Maya (Mexico; P^{IAM}). We begin by combining the HRS and Add Health genotype data with our four reference panels. Next, we restrict to autosomal SNPs that are present in the HRS, Add Health, and the reference panels, leaving 412,491 overlapping SNPs. After implementing linkage disequilibrium pruning (with a window size of 200kb, a step size of 25, and an R^2 of 0.4) in PLINK1.9 (44), we retain 203,429 SNPs for use in GSP estimation. By construction, the four GSP estimates sum to one for each respondent.

$$P_i^{AFR} + P_i^{EUR} + P_i^{EAS} + P_i^{IAM} = 1$$

Eq. 1

Note, while the ADMIXTURE software is intended for use on samples of unrelated individuals, the Add Health data contains a relatively small number of sibling (and half-sibling) pairs. Thus, we remove a random sibling from each pair to create a subsample of unrelated respondents. We first fit supervised ADMIXTURE using only these unrelated respondents, and then project the model results to obtain GSPs for the omitted siblings.

Supplemental unsupervised ADMIXTURE analysis (conducted separately on the HRS and Add Health samples) recovers nearly identical estimates to our supervised ADMIXTURE analysis, suggesting that our GSP results are not sensitive to the specific reference panels utilized; see Figures S4 and S5 for more details. In addition, results from local genetic similarity estimation software (45), when summed across the genome, are highly comparable to those from ADMIXTURE (see Figure S5 from Zhang and Trejo 2025 (25)). Notably, past research using supervised ADMIXTURE has demonstrated that, in U.S. samples, the resulting measures closely correspond to the global information provided by popular genetic ancestry tests (see Figure S14 from Bryc et al. 2015 for a comparison with 23andMe (46)).

Importantly, the process of transforming high-dimensional individual genotype data to a low-dimensional set of individual GSPs is, in effect, an exercise in dimensionality reduction. Any given process for estimating GSPs requires a series of researcher decisions which do not have objectively correct answers, including specifying the number of populations to use, selecting the relevant reference panels, and choosing the algorithm used to cluster and categorize the genome (47). Just as there is no single ‘correct’ way to transform measures of income, education, occupation, and wealth into a ‘true’ measure of socioeconomic status, there is no single correct way to construct

GSPs that yields a ‘true’ measure of genetic ancestry. Nonetheless, we argue that the GSPs used in this study represent a particular manifestation of a set of latent variables which have useful properties for social scientific inquiry.

Racial/Ethnic Identity

Unfortunately, only a limited amount of self-reported racial/ethnic identity information is available in the HRS public survey data. (Note, in order to analyze the HRS and Add Health genotype data jointly on a single server, we were forced to use the publicly available – rather than restricted use – HRS survey data.) HRS respondents were typically asked to report their racial/ethnic identity only at their first interview; in particular, they were asked to select a single primary race from the following five options: ‘White/Caucasian’, ‘Black/African-American’, ‘American Indian or Alaskan Native’, ‘Asian or Pacific Islander’, and ‘Other’.⁵ However, due to concerns regarding respondent identifiability, all racial categories other than White and Black are combined into an ‘Other Races’ category in the public-use survey data. We intersect categorical responses from the racial identity question with binary responses to a Hispanicity question to create the following four mutually exclusive racial categories: Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Other Races, and Hispanic. We leave analysis of the relationship between GSPs and multi-racial identification to future work.

We next sought to construct a self-reported racial/ethnic identity in Add Health that was identical to the HRS variable. Add Health respondents were asked to report their racial identity at Waves I, III, and V. However, only the Wave I survey question matched the survey question used in the HRS. (In particular, the Wave III racial identity question did not contain ‘Other’ as a possible response and, at Wave V, racial identity and Hispanicity were combined into a single question). At Wave I, Add Health respondents were asked to report which one category best describes their racial background from the following five options: ‘White’, ‘Black or African American’, ‘American Indian or Native American’, ‘Asian or Pacific Islander’, and ‘Other’. For analyses with the HRS data, we intersect Wave I racial identity and Hispanicity and then combine ‘American Indian or Native American’, ‘Asian or Pacific Islander’, and ‘Other’ into a single ‘Other Races’ category.⁶ This leaves us with four mutually exclusive racial/ethnic categories: Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Other Races, and Hispanic. Thus, when we refer to ‘Black Americans’ and ‘White Americans,’ we are describing the group of individuals who self-identify as Black or White and who do *not* also identify as Hispanic. Table S4 cross-tabulates respondent racial/ethnic identity at Waves I and V; self-reports of race/ethnicity are highly concordant across the two waves, suggesting that our results are unlikely to be sensitive to the specific Add Health race/ethnicity variable used.

Note that, although early versions of genetic ancestry tests became available to the public around 2008, the boom in popularity did not occur until almost a decade later (48); thus, both HRS and Add Health respondents were asked to report their racial identity *prior* to the widespread proliferation of genetic ancestry tests (and, therefore, such information is highly unlikely to have impacted their racial/ethnic identification).

Inverse Probability Weighting

While over 80% of eligible respondents in both survey samples consented to genotyping, non-random selection into genotyping could undermine our ability to construct analytic samples in HRS and Add Health that are nationally representative of the 1945 and 1980 birth cohorts, respectively.⁷ For this reason, we utilize inverse probability weighting to adjust the raw survey weights for observable differences between the overall HRS and Add Health samples and our analytic subsamples of genotyped respondents (50). Specifically, we fit the following weighted logistic regression model, separately in the HRS and in Add Health:

$$\ln\left(\frac{P(dna_i = 1)}{1 - P(dna_i = 1)}\right) = \beta_0 + \mathbf{X}_i\boldsymbol{\Pi}$$

Eq. 2

⁵ Prior to 2006, HRS respondents were required to select a single race. Starting in 2006, the question format was changed to meet updated U.S. Office of Management and Budget guidelines. Specifically, respondents were now allowed to report multiple races; if a respondents reported more than one race, they were then asked which one they consider to be their primary race. We choose to utilize the primary race variable to maximize comparability of the HRS variable over time.

⁶ Note, our demographic projection model utilizes data from Add Health (but not HRS). Therefore, for that analysis, we do not combine ‘American Indian or Native American’, ‘Asian or Pacific Islander’, and ‘Other’ into a single ‘Other Races’ category.

⁷ See Trejo and Kanopka 2024 (49) for derivations and further discussion of how selection into genotyping can introduce bias into genomic analyses.

where dna_i is a binary variable indicating whether individual i is genotyped and \mathbf{X}_i is a vector of covariates including gender, birth year, racial/ethnic identity, educational attainment, and census region. We then multiply each individual i 's original HRS/Add Health sampling weight by the inverse of their model-derived predicted probability of genotyping (i.e., $\frac{1}{\hat{p}_{dna_i}}$) to produce the final adjusted survey weights used in our analysis. See Table S5 for the results from the logistic regression described in Eq. 2. In both data sets, Black Americans and those of Other Races were less likely to be genotyped than White Americans. In Add Health (but not the HRS), Hispanic Americans were less likely to be genotyped than White Americans. Finally, in both studies, a person's educational attainment and their census region displayed statistically significant associations with their likelihood of being genotyped.

In the Add Health data, ungenotyped respondents typically either attrited from the study before Wave IV or did not consent to genotyping. However, in the HRS, mortality selection may also have played a role. To limit the degree of mortality selection in the HRS, we do not utilize data from birth cohorts prior to 1931; specifically, we omit the Asset and Health Dynamics Among the Oldest Old (1890-1923) and Children of Depression (1924-1930) cohorts. However, mortality selection could still introduce bias into our results if – within self-identified racial/ethnic groups – individuals with certain GSPs were more likely to survive (and therefore remain in the study). To test this possibility, we explore whether our GSPs predict mortality among the genotyped sample. Reassuringly, weighted and unweighted regressions show no statistically significant relationship between any of the GSPs and mortality (net of the inverse probability weighting covariates); see table S6 for more details.

Kitagawa Decomposition

We leverage a Kitagawa decomposition (30) to separate observed average GSP changes across birth cohorts into the sum of two distinct components: [i] the expected change to a given GSP resulting from shifts in racial/ethnic demographics (holding average between-group difference in the GSPs fixed), and [ii] a residual term which corresponds to changes in the GSP within racial/ethnic groups. Specifically, our decomposition takes the following form:

$$\Delta_{45}^{80} \bar{P}^{AFR} = + \sum_{k=1}^K \underbrace{(\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)}_{\text{Racial/Ethnic Change}} + \sum_{k=1}^K \underbrace{(\Delta_{45}^{80} \bar{P}_k^{AFR} \times race_{80,k})}_{\text{Within-Group Change}}$$

where $\Delta_{45}^{80} \bar{P}^{AFR}$ is the overall change in average Sub-Saharan African genetic similarity from the 1945 birth cohort to the 1980 birth cohort, $\bar{P}_{45,k}^{AFR}$ is the average Sub-Saharan African genetic similarity of racial/ethnic group k in 1945, $\Delta_{45}^{80} race_k$ is the change in the population share of racial/ethnic group k the 1945 birth cohort to the 1980 birth cohort, $\Delta_{45}^{80} \bar{P}_k^{AFR}$ is the change in average Sub-Saharan African genetic similarity of racial/ethnic group k from the 1945 birth cohort to the 1980 birth cohort, $race_{80,k}$ is the population share of racial/ethnic group k the 1980 birth cohort, and K is the number of racial/ethnic groups. See Section S1 of the SI for a complete derivation of our specific Kitagawa decomposition.

Notably, while the Kitagawa decomposition is intimately related to the Oaxaca-Blinder decomposition (51, 52), there are important distinctions; indeed, our analysis may be one of the rare cases where the Kitagawa decomposition and Oaxaca-Blinder decompositions meaningfully diverge (53). The Kitagawa decomposition is originally tailored for the decomposition of *proportions* (e.g., subgroup-specific death rates), whereas the Oaxaca-Blinder decomposition can be applied to a wider range of outcome variables; moreover, while the Oaxaca-Blinder decomposition is regression-based (and therefore requires the use of underlying micro data), the Kitagawa decomposition has more flexible data requirements and instead requires only a relevant set of sample moments. Thus, while we are able to utilize a Kitagawa decomposition on the results of the demographic projection for the 2015 birth cohort (described in the following subsection), it would be impossible to apply a Oaxaca-Blinder decomposition (as we lack any genotyped individual-level data of this birth cohort).

Demographic Projection

We utilize a unique demographic projection technique to estimate the average GSPs of the 2015 U.S. birth cohort. In short, we leverage the fact that parents transmit – in expectation – a random 50% of their DNA (and so too, their GSPs) to each of their children. Thus, the expected GSPs of a given birth cohort are simply equal to the average GSPs of the parents of all the members of that birth cohort. In Add Health, respondents were asked to report their number of biological children at Wave V (when they were, on average, about 38 years old). Specifically, women were asked the number of times they have become pregnant, and how many live births resulted from those

pregnancies; men, on the other hand, were asked the number of times a partner has become pregnant, and how many live births resulted from those pregnancies. Respondents were instructed to only include pregnancies in which they were a biological parent. We use this information to calculate the fertility-weighted average value of the GSPs for each sex-racial/ethnic group combination, as follows:

$$\begin{aligned}\tilde{P}_{m,k}^{\text{AFR}} &= \frac{\sum_{i=1}^N (P_i^{\text{AFR}} \times D_i^{\text{sex}=m} \times D_i^{\text{race}=k} \times \text{children}_i)}{\sum_{i=1}^N (D_i^{\text{sex}=m} \times D_i^{\text{race}=k} \times \text{children}_i)} \\ \tilde{P}_{f,k}^{\text{AFR}} &= \frac{\sum_{i=1}^N (P_i^{\text{AFR}} \times D_i^{\text{sex}=f} \times D_i^{\text{race}=k} \times \text{children}_i)}{\sum_{i=1}^N (D_i^{\text{sex}=f} \times D_i^{\text{race}=k} \times \text{children}_i)}\end{aligned}$$

Eq. 3

where P_i^{AFR} is the Sub-Saharan African genetic similarity of individual i , $D_i^{\text{sex}=m}$ is a binary variable indicating that individual i is male, $D_i^{\text{sex}=f}$ is a binary variable indicating that individual i is female, $D_i^{\text{race}=k}$ is a binary variable indicating that individual i self-identifies as race/ethnicity k , children_i is individual i 's number of biological children, and N is the number of genotyped Add Health respondents. Note, while Eq. 1 uses Sub-Saharan African genetic similarity (P^{AFR}) as an example, the same logic applies to the other three GSPs.

Next, we use nationally representative data from the 2013-2017 American Community Survey (ACS) to calculate the racial/ethnic identity population shares of Americans who had a child born in the past year, as follows:

$$\begin{aligned}\text{race}_{m,k} &= \frac{\sum_{i=1}^M (D_i^{\text{race}=k} \times D_i^{\text{sex}=m} \times \text{weight}_i)}{\sum_{i=1}^M (D_i^{\text{sex}=m} \times \text{weight}_i)} \\ \text{race}_{f,k} &= \frac{\sum_{i=1}^M (D_i^{\text{race}=k} \times D_i^{\text{sex}=f} \times \text{weight}_i)}{\sum_{i=1}^M (D_i^{\text{sex}=f} \times \text{weight}_i)}\end{aligned}$$

Eq. 4

where weight is the ACS's individual-level sampling weight and M is the number of recent births in the ACS data.

Finally, we link the fertility-weighted racial/ethnic group GSP averages (calculated in Eq. 3 using Add Health) to the racial/ethnic shares of mothers and fathers Americans who had a child born in the past year (calculated in Eq. 4 using the ACS) to estimate the average GSPs of the 2015 U.S. birth cohort:

$$\hat{P}_{15}^{\text{AFR}} = \frac{1}{2} \times \sum_{k=1}^K (\text{race}_{f,k} \times \tilde{P}_{f,k}^{\text{AFR}} + \text{race}_{m,k} \times \tilde{P}_{m,k}^{\text{AFR}})$$

Eq. 5

where $\hat{P}_{15}^{\text{AFR}}$ is the estimated average Sub-Saharan African genetic similarity of the 2015 birth cohort, $\text{race}_{m,k}$ is the fraction of babies born to fathers who self-identify as racial/ethnic category k , $\text{race}_{f,k}$ is the fraction of babies born to mothers who self-identify as racial/ethnic category k , and $\tilde{P}_{m,k}^{\text{AFR}}$ and $\tilde{P}_{f,k}^{\text{AFR}}$ are the fertility-weighted average Sub-Saharan African genetic similarity for men and women, respectively, of race/ethnicity k . Standard errors, displayed in Table S7, are estimated via the bootstrap through simultaneously resampling observations in both Add Health and the ACS.

The key assumption of our demographic projection model is that fertility-weighted racial/ethnic group GSP averages in Add Health recover the true (unobserved) average GSP value of the parents in the ACS data. One concern is that Add Health respondents were born slightly earlier than the parents of the U.S. 2015 birth cohort. Whereas Add Health respondents have an average birth year of approximately 1980; fathers and mothers in the ACS data have an average birth year of approximately 1982 and 1985, respectively. However, because U.S. birth cohorts are growing more diverse over time, this gap entails that our demographic projection model may slightly *underestimate* the expected change in GSPs from 1980 to 2015. Another concern is that the Add Health fertility data was collected when respondents were around 37 years old, meaning it does not contain children born to parents at older ages; however, in the ACS data, fathers and mothers above the age of 37 accounted for only 19% and 8% of births, respectively – suggesting the resulting biases are likely relatively small. Notably, our projection method does

not assume that the average GSPs of racial/ethnic group remains constant over time; instead, our strategy accounts for within-group changes in average GSPs by leveraging the within-racial/ethnic group variation in fertility that is associated with one's GSPs.

Note, our Kitagawa decomposition requires racial/ethnic identity shares for the 2015 birth cohort; however, we do not observe child racial/ethnic identity in the ACS (and, indeed, this cohort of children is still relatively young and therefore may not have formed a clear racial/ethnic identity). To address this challenge, we utilized the fact that Add Health has information on both child and parental racial/ethnic identity. Thus, we simply use Add Health to create a copula that contains a child's probability of identifying as a member of each racial/ethnic category given the racial/ethnic identity of their mother and racial/ethnic identity of their father. We then apply this copula to the parental racial/ethnic variables in the ACS data, giving us our final 2015 child racial/ethnic shares.

References and Notes

1. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglu, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kasprzyk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M. J. Morgan, International Human Genome Sequencing Consortium, C. for G. R. Whitehead Institute for Biomedical Research, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, I. of M. B. Department of Genome Analysis, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, T. I. for S. B. Multimegabase Sequencing Center, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, L. A. H. G. C. Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology:, also includes individuals listed under other headings): *Genome Analysis Group (listed in alphabetical order, U. N. I. of H. Scientific management: National Human Genome Research Institute, Stanford Human Genome Center:, University of Washington Genome Center:, K. U. S. of M. Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas:, U. D. of E. Office of Science, The Wellcome Trust:, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. A. Abdellaoui, L. Yengo, K. J. H. Verweij, P. M. Visscher, 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics* **110**, 179–194 (2023).
3. E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, D. Posthuma, Genome-wide association studies. *Nat Rev Methods Primers* **1**, 59 (2021).
4. D. Reich, *Who We Are and How We Got Here: Ancient DNA and the New Science of the Human Past* (Oxford University Press, 2018).
5. L. Orlando, R. Allaby, P. Skoglund, C. Der Sarkissian, P. W. Stockhammer, M. C. Ávila-Arcos, Q. Fu, J. Krause, E. Willerslev, A. C. Stone, C. Warinner, Ancient DNA analysis. *Nat Rev Methods Primers* **1**, 14 (2021).

6. R. Nielsen, J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff, E. Willerslev, Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
7. G. Livingston, A. Brown, “Intermarriage in the U.S. 50 Years After Loving v. Virginia” (Pew Research Center, 2017); <https://www.pewresearch.org/social-trends/2017/05/18/intermarriage-in-the-u-s-50-years-after-loving-v-virginia/>.
8. M. Hout, J. R. Goldstein, How 4.5 Million Irish Immigrants Became 40 Million Irish Americans: Demographic and Subjective Aspects of the Ethnic Composition of White Americans. *American Sociological Review* **59**, 64 (1994).
9. R. D. Flores, M. V. Loría, R. M. Casas, Transitory versus Durable Boundary Crossing: What Explains the Indigenous Population Boom in Mexico? *American Journal of Sociology* **129**, 123–161 (2023).
10. J. Lee, F. D. Bean, America’s Changing Color Lines: Immigration, Race/Ethnicity, and Multiracial Identification. *Annu. Rev. Sociol.* **30**, 221–242 (2004).
11. F. J. Davis, *Who Is Black? One Nation’s Definition* (Pennsylvania State University Press, University Park, Pa, 10th anniversary ed., 2001).
12. M. Nobles, *Shades of Citizenship: Race and the Census in Modern Politics* (Stanford University Press, Stanford, CA, 2000).
13. I. Mathieson, A. Scally, What is ancestry? *PLoS Genet* **16**, e1008624 (2020).
14. A. Sonnega, J. D. Faul, M. B. Ofstedal, K. M. Langa, J. W. Phillips, D. R. Weir, Cohort Profile: the Health and Retirement Study (HRS). *International Journal of Epidemiology* **43**, 576–585 (2014).
15. K. M. Harris, C. T. Halpern, E. A. Whitsel, J. M. Hussey, L. A. Killeya-Jones, J. Tabor, S. C. Dean, Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health). *International Journal of Epidemiology* **48**, 1415–1415k (2019).
16. National Academies of Sciences, Engineering, and Medicine, *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* (National Academies Press, Washington, D.C., 2023; <https://www.nap.edu/catalog/26902>).
17. I. H. 3 Consortium, Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).
18. A. Bergström, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, others, Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
19. G. Coop, Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics. arXiv [Preprint] (2022). <https://doi.org/10.48550/ARXIV.2207.11595>.
20. Y. Ding, K. Hou, Z. Xu, A. Pimplaskar, E. Petter, K. Boulier, F. Privé, B. J. Vilhjálmsson, L. M. Olde Loohuis, B. Pasaniuc, Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).
21. W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. A. Rojo Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, D. Reich, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).

22. G. L. Wojcik, M. Graff, K. K. Nishimura, R. Tao, J. Haessler, C. R. Gignoux, H. M. Highland, Y. M. Patel, E. P. Sorokin, C. L. Avery, G. M. Belbin, S. A. Bien, I. Cheng, S. Cullina, C. J. Hodonsky, Y. Hu, L. M. Huckins, J. Jeff, A. E. Justice, J. M. Kocarnik, U. Lim, B. M. Lin, Y. Lu, S. C. Nelson, S.-S. L. Park, H. Poisner, M. H. Preuss, M. A. Richard, C. Schurmann, V. W. Setiawan, A. Sockell, K. Vahi, M. Verbanck, A. Vishnu, R. W. Walker, K. L. Young, N. Zubair, V. Acuña-Alonso, J. L. Ambite, K. C. Barnes, E. Boerwinkle, E. P. Bottinger, C. D. Bustamante, C. Caberto, S. Canizales-Quinteros, M. P. Conomos, E. Deelman, R. Do, K. Doheny, L. Fernández-Rhodes, M. Fornage, B. Hailu, G. Heiss, B. M. Henn, L. A. Hindorff, R. D. Jackson, C. A. Laurie, C. C. Laurie, Y. Li, D.-Y. Lin, A. Moreno-Estrada, G. Nadkarni, P. J. Norman, L. C. Pooler, A. P. Reiner, J. Romm, C. Sabatti, K. Sandoval, X. Sheng, E. A. Stahl, D. O. Stram, T. A. Thornton, C. L. Wassel, L. R. Wilkens, C. A. Winkler, S. Yoneyama, S. Buyske, C. A. Haiman, C. Kooperberg, L. Le Marchand, R. J. F. Loos, T. C. Matise, K. E. North, U. Peters, E. E. Kenny, C. S. Carlson, Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
23. R. A. Patel, S. A. Musharoff, J. P. Spence, H. Pimentel, C. Tcheandjieu, H. Mostafavi, N. Sinnott-Armstrong, S. L. Clarke, C. J. Smith, P. P. Durda, K. D. Taylor, R. Tracy, Y. Liu, W. C. Johnson, F. Aguet, K. G. Ardlie, S. Gabriel, J. Smith, D. A. Nickerson, S. S. Rich, J. I. Rotter, P. S. Tsao, T. L. Assimes, J. K. Pritchard, Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *The American Journal of Human Genetics* **109**, 1286–1297 (2022).
24. G. Guo, Y. Fu, H. Lee, T. Cai, K. Mullan Harris, Y. Li, Genetic Bio-Ancestry and Social Construction of Racial Classification in Social Surveys in the Contemporary United States. *Demography* **51**, 141–172 (2014).
25. L. Zhang, S. Trejo, DNA, Self-Reported Ancestry, and Social Scientific Inquiry. *SocArXiv*, doi: 10.31235/osf.io/mdybz_v1 (2025).
26. B. Taddess, S. Trejo, L. Zhang, Leveraging Genomic Data to Document Within-Race Attractiveness Penalties Among Black Americans. *SocArXiv* (2025).
27. B. Duncan, S. J. Trejo, Intermarriage and the intergenerational transmission of ethnic identity and human capital for Mexican Americans. *Journal of labor economics* **29**, 195–227 (2011).
28. W. D. Roth, E. G. van Stee, A. Regla-Vargas, Conceptualizations of Race: Essentialism and Constructivism. *Annu. Rev. Sociol.* **49**, annurev-soc-031021-034017 (2023).
29. S. Ruggles, S. Flood, M. Sobek, D. Backman, G. Cooper, J. A. R. Drew, S. Richards, R. Rogers, J. Schroeder, K. C. W. Williams, IPUMS USA: Version 16.0, IPUMS (2025); <https://doi.org/10.18128/D010.V16.0>.
30. E. M. Kitagawa, Components of a difference Between two rates. *Journal of the American Statistical Association* **50**, 1168–1194 (1955).
31. A. Morning, *The Nature of Race: How Scientists Think and Teach about Human Difference* (2011).
32. D. E. Roberts, *Fatal Invention: How Science, Politics, and Big Business Re-Crete Race in the Twenty-First Century* (The New Press, New York, Paperback edition., 2011).
33. A. Nelson, *The Social Life of DNA: Race, Reparations, and Reconciliation after the Genome* (Beacon Press, Boston, 2016).
34. S. Trejo, D. O. Martschenko, *What We Inherit: How New Technologies and Old Myths Are Shaping Our Genomic Future* (Princeton University Press, 2026).
35. G. R. Cutting, Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat Rev Genet* **16**, 45–56 (2015).

36. F. B. Piel, A. P. Patil, R. E. Howes, O. A. Nyangiri, P. W. Gething, T. N. Williams, D. J. Weatherall, S. I. Hay, Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat Commun* **1**, 104 (2010).
37. M. C. Mills, C. Rahal, The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat Genet* **52**, 242–243 (2020).
38. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, M. J. Daly, Current clinical use of polygenic scores will risk exacerbating health disparities. *Nature genetics* **51**, 584 (2019).
39. L. J. Beesley, M. Salvatore, L. G. Fritsche, A. Pandit, A. Rao, C. Brummett, C. J. Willer, L. D. Lisabeth, B. Mukherjee, The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Statistics in medicine* **39**, 773–800 (2020).
40. S. Benonisdottir, A. Kong, Studying the genetics of participation using footprints left on the ascertained genotypes. *Nature Genetics* **55**, 1413–1420 (2023).
41. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655–1664 (2009).
42. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 1–6 (2011).
43. S. S. Shringarpure, C. D. Bustamante, K. Lange, D. H. Alexander, Efficient analysis of large datasets and sex bias with ADMIXTURE. *BMC Bioinformatics* **17**, 218 (2016).
44. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).
45. H. Hilmarsson, A. S. Kumar, R. Rastogi, C. D. Bustamante, D. M. Montserrat, A. G. Ioannidis, High Resolution Ancestry Deconvolution for Next Generation Genomic Data. bioRxiv [Preprint] (2021). <https://doi.org/10.1101/2021.09.19.460980>.
46. K. Bryc, E. Y. Durand, J. M. Macpherson, D. Reich, J. L. Mountain, The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics* **96**, 37–53 (2015).
47. C. D. Royal, J. Novembre, S. M. Fullerton, D. B. Goldstein, J. C. Long, M. J. Bamshad, A. G. Clark, Inferring Genetic Ancestry: Opportunities, Challenges, and Implications. *The American Journal of Human Genetics* **86**, 661–673 (2010).
48. A. Regalado, More than 26 million people have taken an at-home ancestry test, *MIT Technology Review* (2019).
49. S. Trejo, K. Kanopka, Using the phenotype differences model to identify genetic effects in samples of partially genotyped sibling pairs. *Proceedings of the National Academy of Sciences* **121**, e2405725121 (2024).
50. B. W. Domingue, D. W. Belsky, A. Harrati, D. Conley, D. R. Weir, J. D. Boardman, Mortality selection in a genetic sample and implications for association studies. *International journal of epidemiology* **46**, 1285–1294 (2017).
51. A. S. Blinder, Wage Discrimination: Reduced Form and Structural Estimates. *The Journal of Human Resources* **8**, 436–455 (1973).

52. R. Oaxaca, Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review* **14**, 693–709 (1973).
53. R. L. Oaxaca, E. Sierminska, Oaxaca-Blinder meets Kitagawa: What is the link? *PLOS ONE* **20**, e0321874 (2025).

Acknowledgments: We are grateful to Dalton Conley, Augustine Kong, Melinda Mills, and Luyin Zhang for helpful comments. This research uses data from Add Health, funded by grant P01 HD31921 (Harris) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Add Health is currently directed by Robert A. Hummer and funded by the National Institute on Aging cooperative agreements U01 AG071448 (Hummer) and U01AG071450 (Hummer and Aiello) at the University of North Carolina at Chapel Hill. Add Health was designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill. No direct support was received from grant P01 HD31921 for this analysis. Information on obtaining Add Health data is available on the project website.

Funding: This work has been supported by a grant from the Princeton Data Driven Social Sciences Initiative.

Author contributions: ST designed the research. ST and MET analyzed the data. ST and MET wrote the paper.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All results needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Our analyses utilize the RAND HRS Longitudinal File 1992-2020 (V1), which is publicly available and can be downloaded from the following website: <https://hrsdata.isr.umich.edu/data-products/rand>. We also used the restricted HRS Genotype Data Version 2 (2006-2010 Samples), which can be accessed by researchers via application at <https://dbgap.ncbi.nlm.nih.gov>. Finally, we utilized the restricted Add Health survey and genotype data, which can be accessed by researchers via application at <https://data.cpc.unc.edu/projects/2/view>.

Supplementary Materials

Figures S1 to S5

Tables S1 to S7

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Cohort Averages			Kitagawa Decomposition					
Self-Reported Race/Ethnicity	<u>1945</u>	<u>1980</u>	<u>2015</u>	<u>1945 to 1980</u>			<u>1945 to 2015</u>		
				<i>Percent Change</i>			<i>Percent Change</i>		
Non-Hispanic White	0.833	0.714	0.526	-14.2%			-36.8%		
Non-Hispanic Black	0.109	0.153	0.152	40.3%			39.8%		
Hispanic	0.039	0.099	0.242	158.1%			528.7%		
Non-Hispanic Other	0.02	0.033	0.079	69.3%			302.9%		
Genetic Similarity Proportions	<u>1945</u>	<u>1980</u>	<u>2015</u>	<u>1945 to 1980</u>			<u>1945 to 2015</u>		
				<i>Percent Change</i>	<i>Demographic</i>	<i>Within-Group</i>	<i>Percent Change</i>	<i>Demographic</i>	<i>Within-Group</i>
p^{AFR}	0.101	0.138	0.146	36.4%	40.3%	-3.9%	44.4%	52.0%	-7.6%
p^{EUR}	0.864	0.796	0.717	-7.8%	-8.0%	0.2%	-17.0%	-17.6%	0.6%
p^{EAS}	0.011	0.029	0.062	167.9%	23.7%	144.2%	480.3%	97.8%	382.5%
p^{IAM}	0.025	0.038	0.075	52.0%	106.4%	-54.4%	203.4%	360.2%	-156.7%

Tab. 1. Demographic Projection and Kitagawa Decomposition.

This table displays average genetic similarity proportions in three different American birth cohorts: 1945, 1980, and 2015. Column 1 and Column 2 display (weighted) average racial/ethnic shares and genetic similarity proportions of U.S.-born respondents from the Health and Retirement Study and the Add Health Study, respectively. Column 3 displays estimated racial/ethnic shares and genetic similarity proportions obtained via our demographic projection model. Columns 4 through 9 present results from a Kitagawa decomposition, which separates changes to a given genetic similarity proportion over time into the expected change given shifts in racial/ethnic demographics and a residual term corresponding to changes in the average genetic similarity proportions within racial/ethnic groups. Survey weights in both analytic samples are adjusted for selection into genotyping using inverse probability weighting. Genetic similarity proportions are estimated via supervised ADMIXTURE and correspond to Sub-Saharan African (p^{AFR}), European (p^{EUR}), East Asian (p^{EAS}), and Indigenous American (p^{IAM}) genetic similarity. See Table S7 for an alternative version of this table that includes standard errors, which are generally quite small.

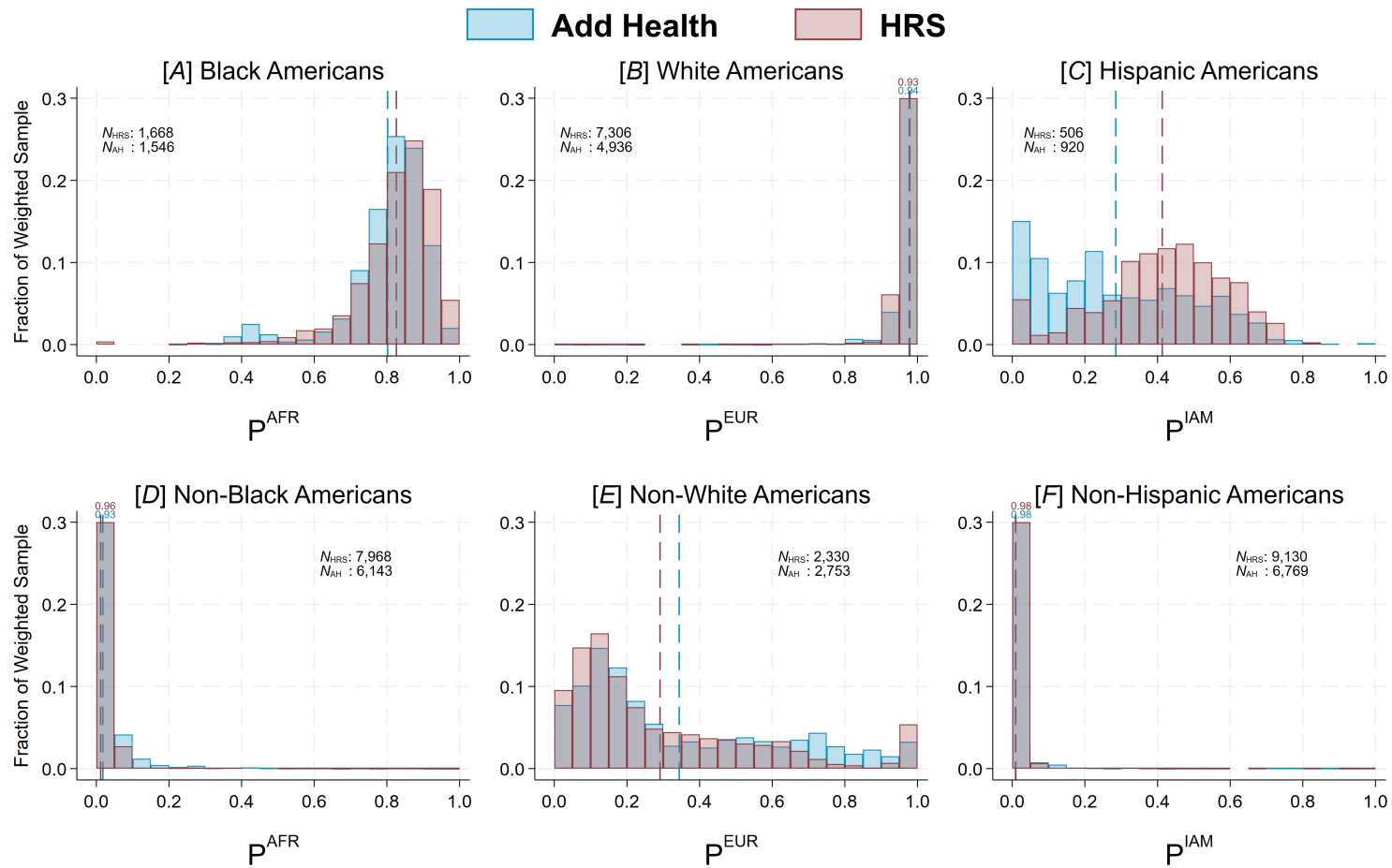


Fig. 1. Within-race/ethnicity changes in genetic similarity proportions from 1945 to 1980.

This figure contains histograms of the distribution of the Sub-Saharan African (P^{AFR} ; Panels A and D), European (P^{EUR} ; Panels B and E), and Indigenous American (P^{IAM} ; Panels C and F) genetic similarity proportions for various racial/ethnic groups. Genetic similarity proportions are estimated via supervised ADMIXTURE. Racial/ethnic categories (Non-Hispanic Black, Non-Hispanic White, and Hispanic) are mutually exclusive. The red bars display data from the Health and Retirement Study, which has an average birth year of approximately 1945; the blue bars display data from the Add Health Study, which has an average birth year of roughly 1980. Only U.S.-born individuals are displayed. Survey weights in both analytic samples are adjusted for selection into genotyping using inverse probability weighting. Bars that would extend beyond the Y-axis values of each histogram are censored, with the true height listed in the corresponding color above the bar.

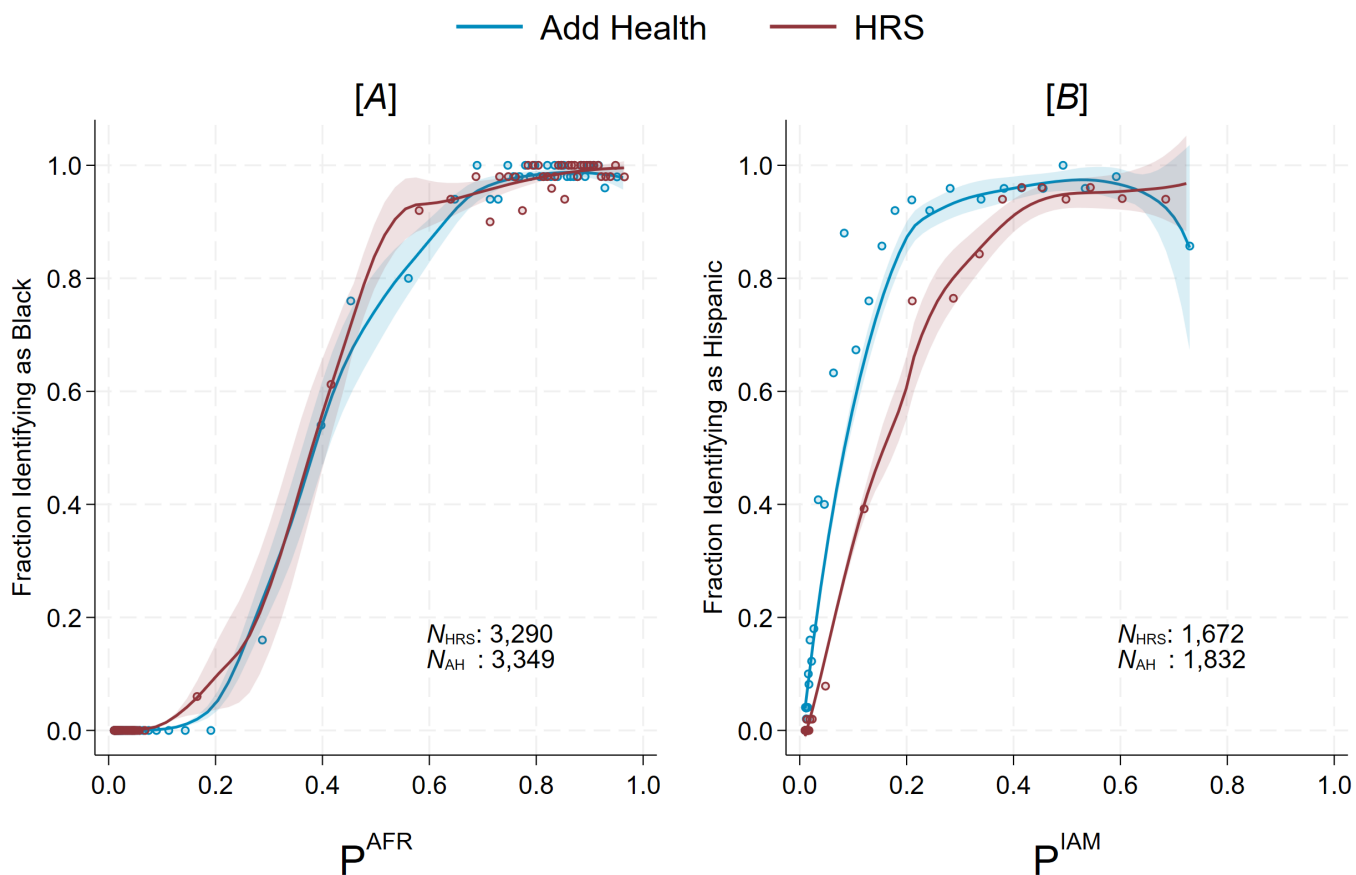


Fig. 2. The association between genetic similarity proportions and racial/ethnic identification in 1945 and 1980.

This figure contains binned scatter plots and local polynomial fit lines (including 95% confidence intervals), with genetic similarity proportions on the X-axis and racial/ethnic identification on the Y-axis. Panel A displays the association between Sub-Saharan African genetic similarity (P^{AFR}) and Non-Hispanic Black identification, whereas the right panel displays the association between Indigenous American genetic similarity (P^{IAM}) and Hispanic identification. Genetic similarity proportions are estimated via supervised ADMIXTURE. In both panels, the red markers and lines display data from the Health and Retirement Study (which has an average birth year of approximately 1945), whereas the blue markers and lines display data from the Add Health Study (which has an average birth year of roughly 1980). Only U.S. born individuals whose value for both the European genetic similarity proportion (P^{EUR}) and the genetic similarity proportion displayed on the X-axis are both greater than 0.01 are included. In addition, individuals whose other two genetic similarity proportions total to greater than 0.5 are excluded. Bins in the left panel contain approximately 65 respondents, and bins in the right panel contain approximately 35 respondents.

SUPPLEMENTARY INFORMATION

Trejo and Thompson 2025

S1. Kitagawa Decomposition Derivation

We begin our derivation with the following tautology:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \bar{P}_{80}^{AFR} - \bar{P}_{45}^{AFR}$$

Next, we replace the overall population GSP average for the 1980 and 1945 births cohorts as the weighted-average of the racial/ethnic group-specific GSP averages, leaving us with:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K (\bar{P}_{80,k}^{AFR} \times race_{80,k}) - \sum_{k=1}^K (\bar{P}_{45,k}^{AFR} \times race_{45,k})$$

Now, noticing that $race_{45,k} = race_{80,k} - \Delta_{45}^{80} race_k$, we swap in as follows:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K (\bar{P}_{80,k}^{AFR} \times race_{80,k}) - \sum_{k=1}^K (\bar{P}_{45,k}^{AFR} \times [race_{80,k} - \Delta_{45}^{80} race_k])$$

Distributing the $\bar{P}_{45,k}^{AFR}$ term in the right-hand summation operator, we have:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K (\bar{P}_{80,k}^{AFR} \times race_{80,k}) - \sum_{k=1}^K [(\bar{P}_{45,k}^{AFR} \times race_{80,k}) - (\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)]$$

Next, we combine everything into a single summation operator and are left with:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K [(\bar{P}_{80,k}^{AFR} \times race_{80,k}) - (\bar{P}_{45,k}^{AFR} \times race_{80,k}) + (\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)]$$

We then rearrange and separate into two summation operators:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K [(\bar{P}_{80,k}^{AFR} \times race_{80,k}) - (\bar{P}_{45,k}^{AFR} \times race_{80,k})] + \sum_{k=1}^K (\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)$$

Noticing that $\Delta_{45}^{80} \bar{P}^{AFR} = \bar{P}_{80,k}^{AFR} - \bar{P}_{45,k}^{AFR}$, we swap in as follows:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K (\Delta_{45}^{80} \bar{P}^{AFR} \times race_{80,k}) + \sum_{k=1}^K (\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)$$

Rearranging, we are left with our final decomposition formulation:

$$\Delta_{45}^{80} \bar{P}^{AFR} = + \sum_{k=1}^K \underbrace{(\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)}_{\text{Racial/Ethnic Change}} + \sum_{k=1}^K \underbrace{(\Delta_{45}^{80} \bar{P}_k^{AFR} \times race_{80,k})}_{\text{Within-Group Change}}$$

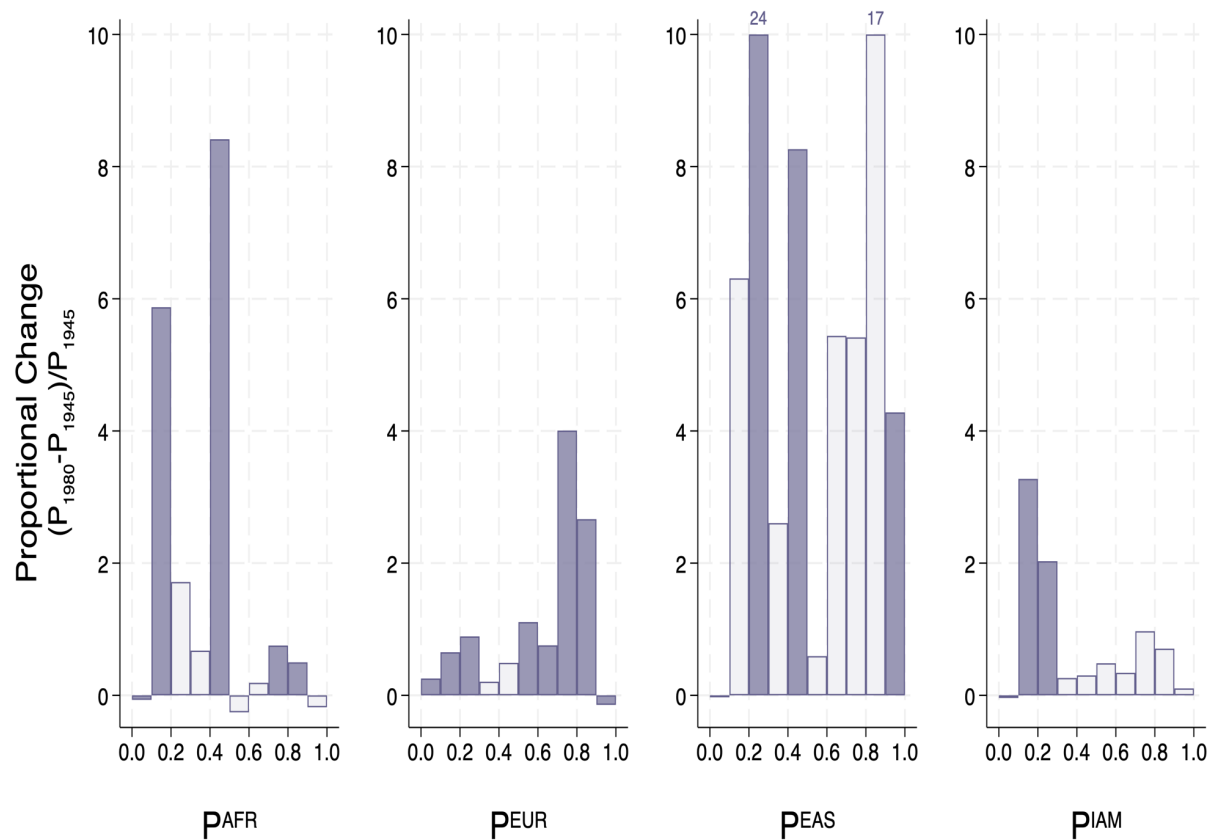


Fig. S1. Proportional change in genetic similarity proportion deciles from 1945 to 1980.

This figure displays the proportional change of p^{AFR} , p^{EUR} , p^{EAS} , and p^{IAM} deciles from the Health and Retirement Study (which has an average birth year of approximately 1945) to the Add Health Study (which has an average birth year of roughly 1980). Dark-colored bars indicate a statistically significant change in a given genetic similarity proportion between cohorts, whereas light-colored bars indicate a change that is not statistically significant ($p > 0.05$). Genetic similarity proportions are estimated via supervised ADMIXTURE. Only U.S.-born individuals are displayed. Survey weights in both analytic samples are adjusted for selection into genotyping using inverse probability weighting. Bars that would extend beyond the Y-axis values of each histogram are censored, with the true height listed above the bar.

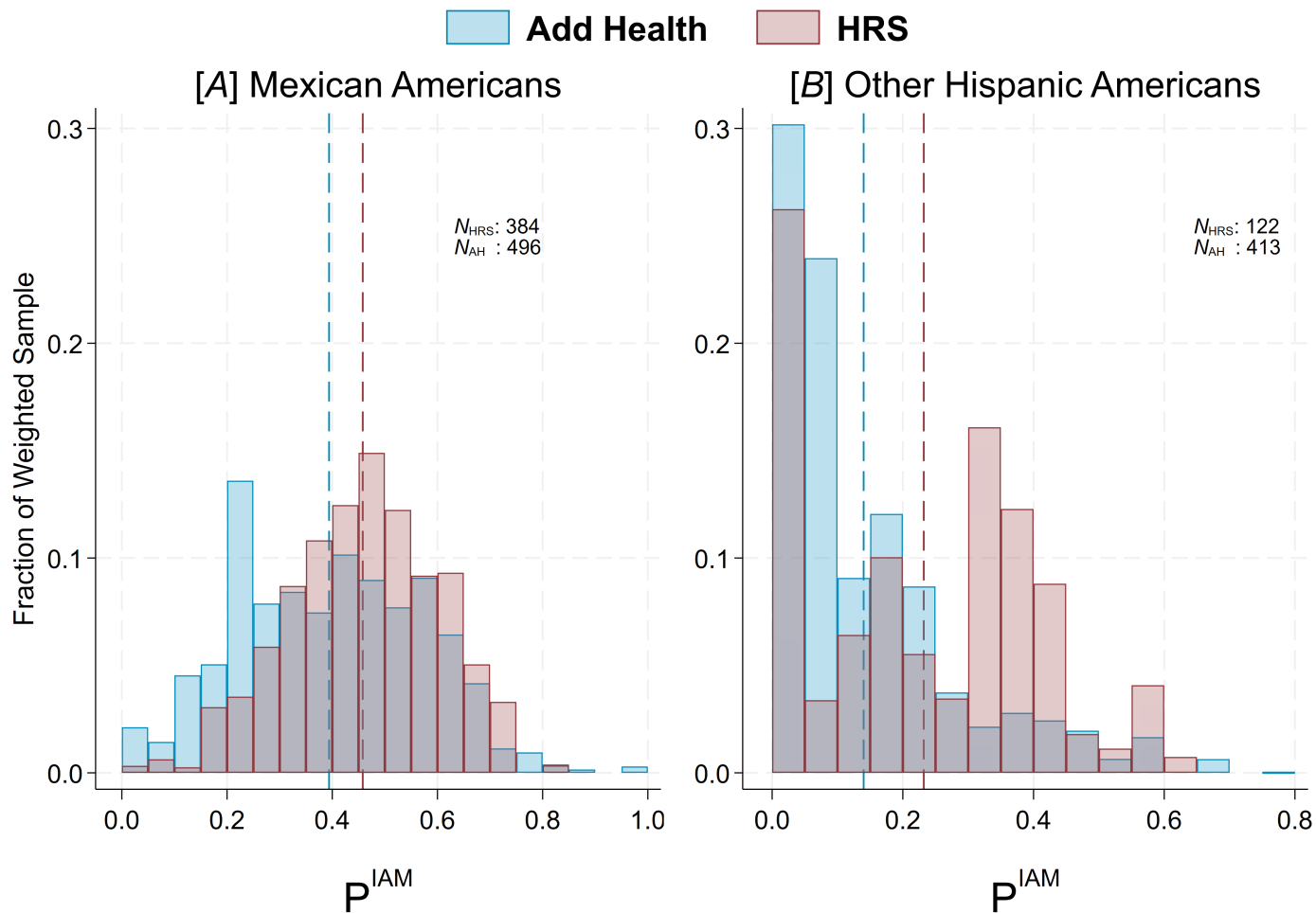


Fig. S2. Changes in Indigenous American genetic similarity from 1945 to 1980 among Hispanic Americans.

This figure contains histograms of the distribution of the Indigenous American genetic similarity proportion (P^{IAM}) separately for Mexican Americans and for Other Hispanic Americans. Genetic similarity proportions are estimated via supervised ADMIXTURE. The red bars display data from the Health and Retirement Study, which has an average birth year of approximately 1945; the blue bars display data from the Add Health Study, which has an average birth year of roughly 1980. Only U.S.-born individuals are displayed. Survey weights in both analytic samples are adjusted for selection into genotyping using inverse probability weighting.

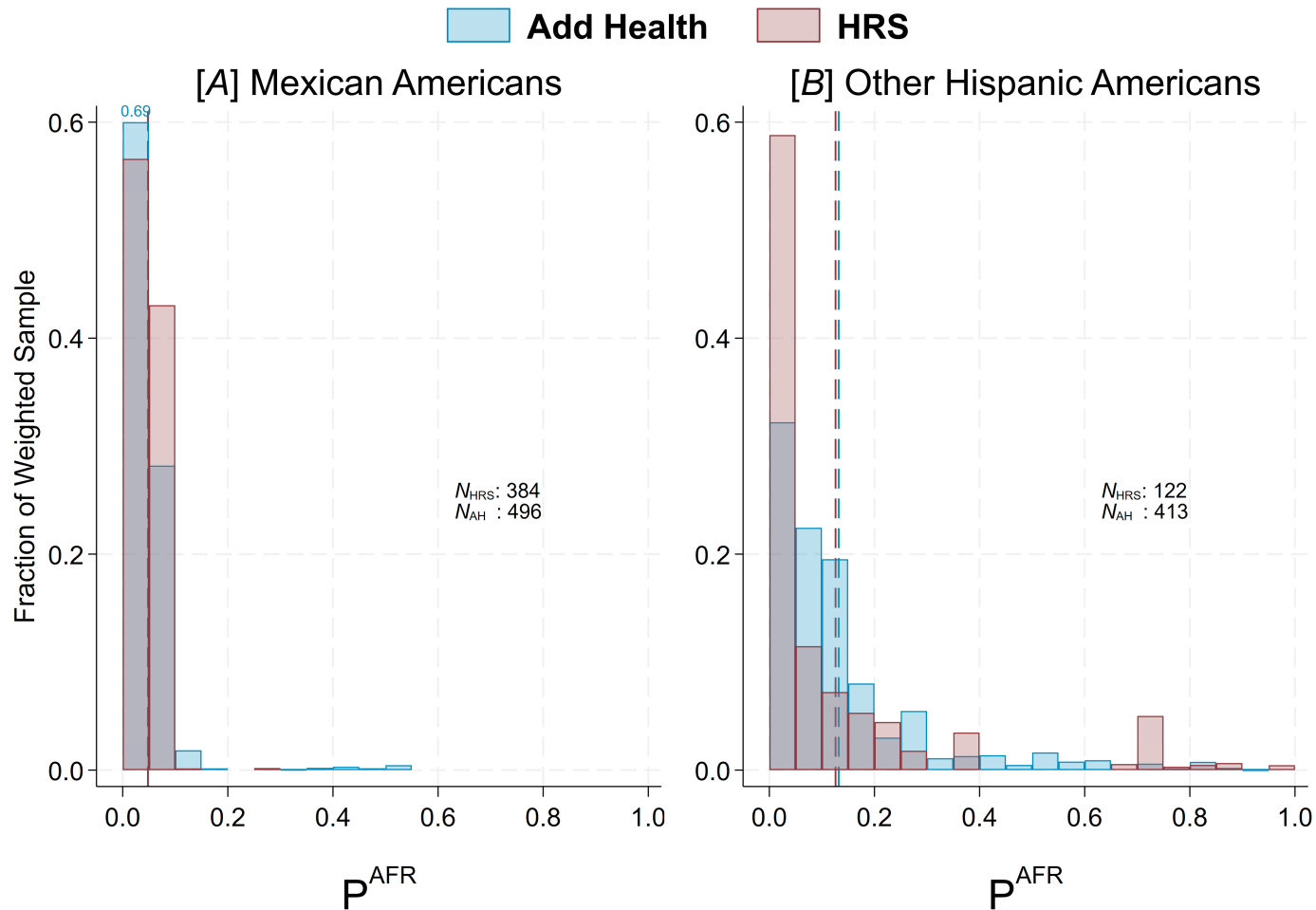


Fig. S3. Changes in Sub-Saharan African genetic similarity from 1945 to 1980 among Hispanic Americans.

This figure contains histograms of the distribution of the Sub-Saharan African genetic similarity proportion (P^{AFR}) separately for Mexican Americans and for Other Hispanic Americans. Genetic similarity proportions are estimated via supervised ADMIXTURE. The red bars display data from the Health and Retirement Study, which has an average birth year of approximately 1945; the blue bars display data from the Add Health Study, which has an average birth year of roughly 1980. Only U.S.-born individuals are displayed. Survey weights in both analytic samples are adjusted for selection into genotyping using inverse probability weighting.

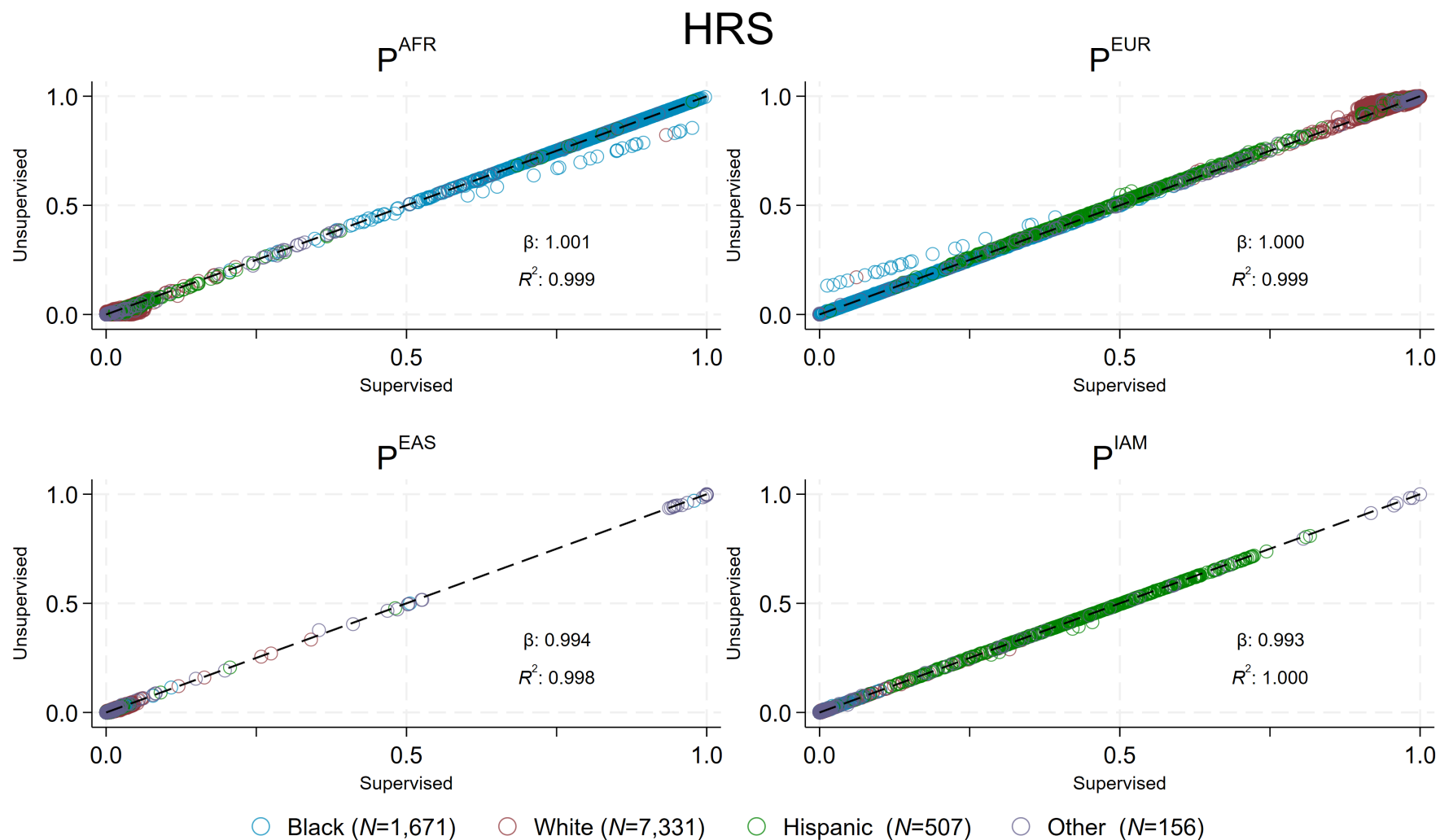


Fig. S4. Correspondence between supervised and unsupervised ADMIXTURE estimates in the Health and Retirement Study.

This figure uses scatter plots to compare the genetic similarity proportions of HRS respondents from supervised and unsupervised ADMIXTURE estimation ($K = 4$). Unsupervised ADMIXTURE estimates are derived using only the HRS sample. Markers are color-coded by respondent racial/ethnic identity.

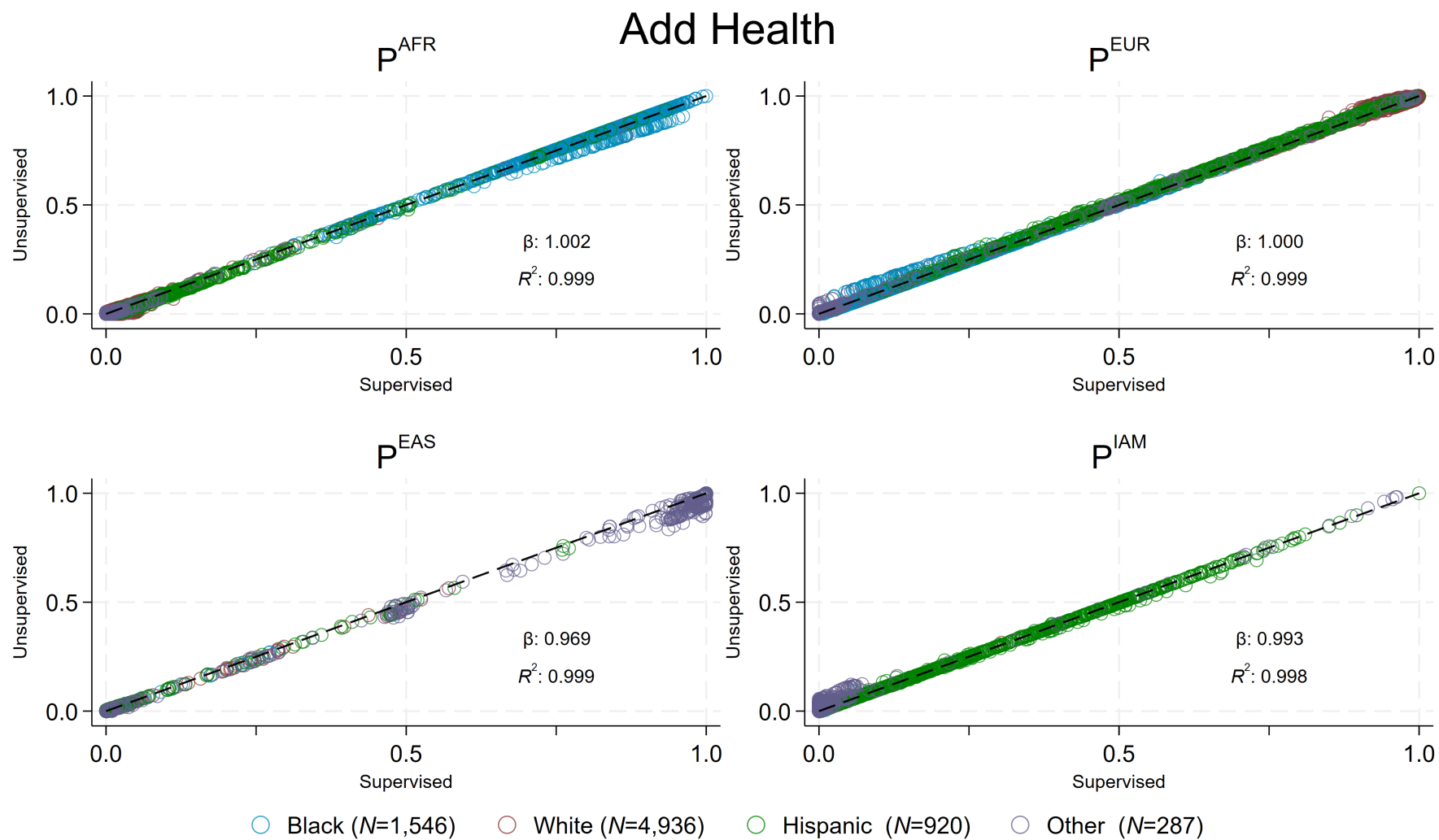


Fig. S5. Correspondence between supervised and unsupervised ADMIXTURE estimates in the Add Health study.

This figure uses scatter plots to compare the genetic similarity proportions of Add Health respondents from supervised and unsupervised ADMIXTURE estimation ($K = 4$). Unsupervised ADMIXTURE estimates are derived using only the Add Health sample. Markers are color-coded by respondent racial/ethnic identity.

SUPPLEMENTARY INFORMATION

Trejo and Thompson 2025

S1. Kitagawa Decomposition Derivation

We begin our derivation with the following tautology:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \bar{P}_{80}^{AFR} - \bar{P}_{45}^{AFR}$$

Next, we replace the overall population GSP average for the 1980 and 1945 births cohorts as the weighted-average of the racial/ethnic group-specific GSP averages, leaving us with:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K (\bar{P}_{80,k}^{AFR} \times race_{80,k}) - \sum_{k=1}^K (\bar{P}_{45,k}^{AFR} \times race_{45,k})$$

Now, noticing that $race_{45,k} = race_{80,k} - \Delta_{45}^{80} race_k$, we swap in as follows:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K (\bar{P}_{80,k}^{AFR} \times race_{80,k}) - \sum_{k=1}^K (\bar{P}_{45,k}^{AFR} \times [race_{80,k} - \Delta_{45}^{80} race_k])$$

Distributing the $\bar{P}_{45,k}^{AFR}$ term in the right-hand summation operator, we have:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K (\bar{P}_{80,k}^{AFR} \times race_{80,k}) - \sum_{k=1}^K [(\bar{P}_{45,k}^{AFR} \times race_{80,k}) - (\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)]$$

Next, we combine everything into a single summation operator and are left with:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K [(\bar{P}_{80,k}^{AFR} \times race_{80,k}) - (\bar{P}_{45,k}^{AFR} \times race_{80,k}) + (\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)]$$

We then rearrange and separate into two summation operators:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K [(\bar{P}_{80,k}^{AFR} \times race_{80,k}) - (\bar{P}_{45,k}^{AFR} \times race_{80,k})] + \sum_{k=1}^K (\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)$$

Noticing that $\Delta_{45}^{80} \bar{P}^{AFR} = \bar{P}_{80,k}^{AFR} - \bar{P}_{45,k}^{AFR}$, we swap in as follows:

$$\Delta_{45}^{80} \bar{P}^{AFR} = \sum_{k=1}^K (\Delta_{45}^{80} \bar{P}^{AFR} \times race_{80,k}) + \sum_{k=1}^K (\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)$$

Rearranging, we are left with our final decomposition formulation:

$$\Delta_{45}^{80} \bar{P}^{AFR} = + \sum_{k=1}^K \underbrace{(\bar{P}_{45,k}^{AFR} \times \Delta_{45}^{80} race_k)}_{\text{Racial/Ethnic Change}} + \sum_{k=1}^K \underbrace{(\Delta_{45}^{80} \bar{P}_k^{AFR} \times race_{80,k})}_{\text{Within-Group Change}}$$

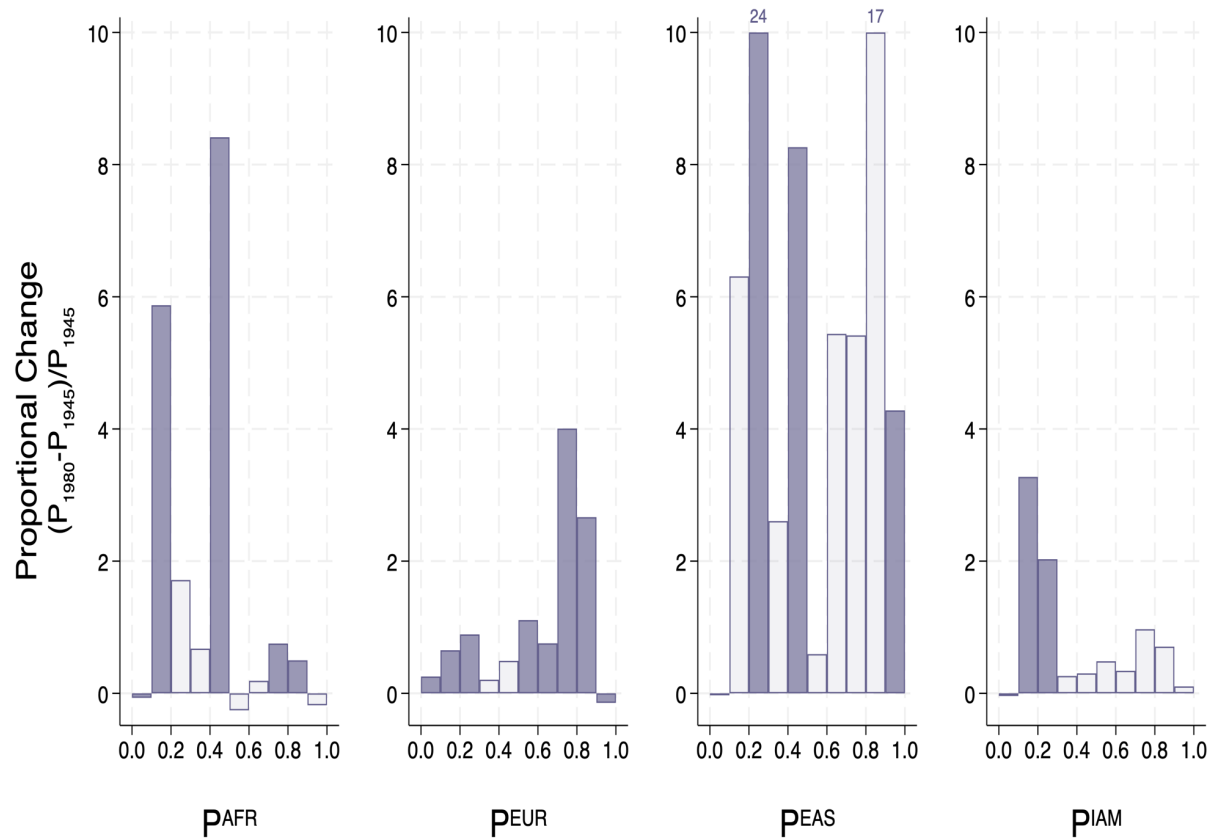


Fig. S1. Proportional change in genetic similarity proportion deciles from 1945 to 1980.

This figure displays the proportional change of p^{AFR} , p^{EUR} , p^{EAS} , and p^{IAM} deciles from the Health and Retirement Study (which has an average birth year of approximately 1945) to the Add Health Study (which has an average birth year of roughly 1980). Dark-colored bars indicate a statistically significant change in a given genetic similarity proportion between cohorts, whereas light-colored bars indicate a change that is not statistically significant ($p > 0.05$). Genetic similarity proportions are estimated via supervised ADMIXTURE. Only U.S.-born individuals are displayed. Survey weights in both analytic samples are adjusted for selection into genotyping using inverse probability weighting. Bars that would extend beyond the Y-axis values of each histogram are censored, with the true height listed above the bar.

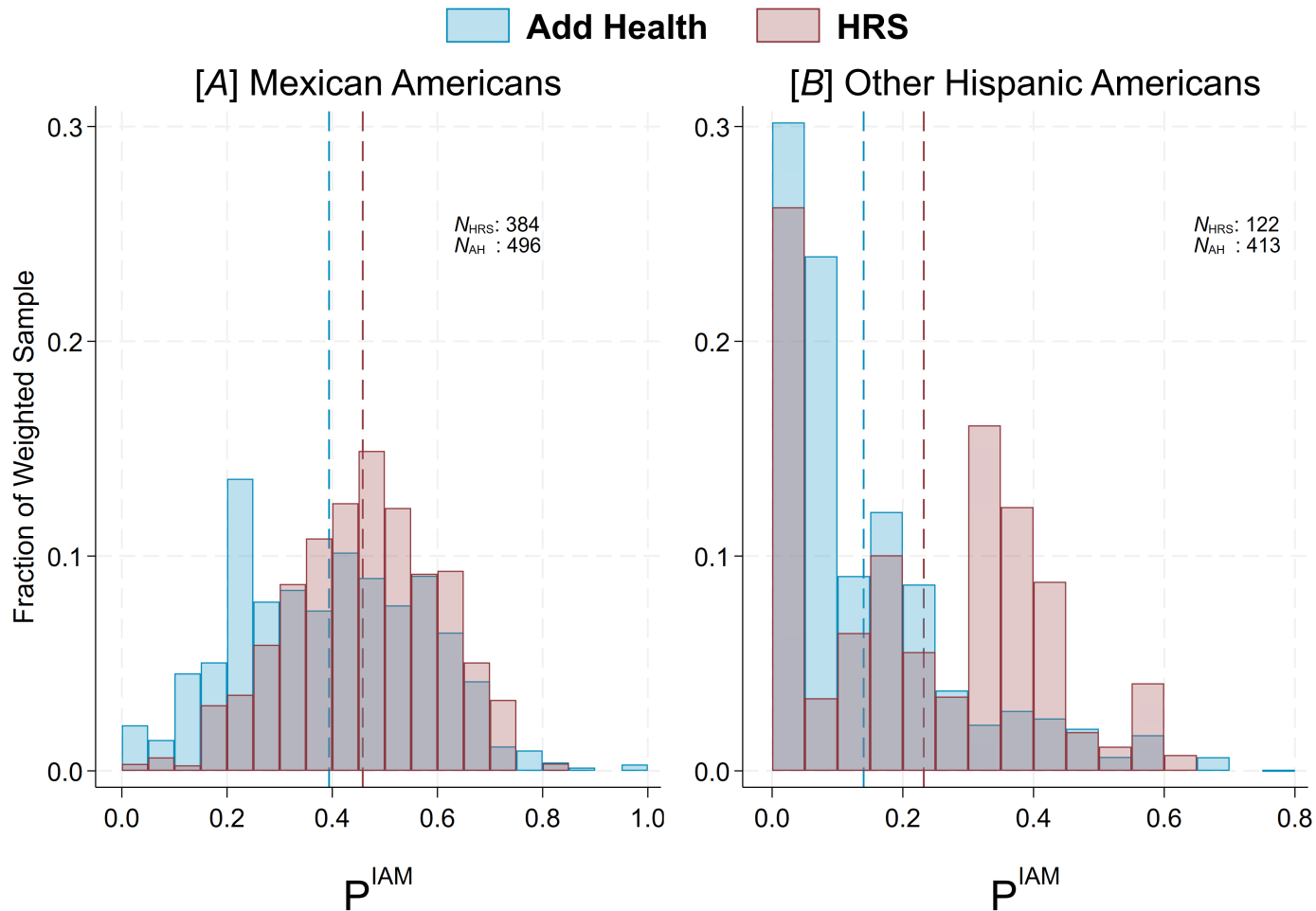


Fig. S2. Changes in Indigenous American genetic similarity from 1945 to 1980 among Hispanic Americans.

This figure contains histograms of the distribution of the Indigenous American genetic similarity proportion (P^{IAM}) separately for Mexican Americans and for Other Hispanic Americans. Genetic similarity proportions are estimated via supervised ADMIXTURE. The red bars display data from the Health and Retirement Study, which has an average birth year of approximately 1945; the blue bars display data from the Add Health Study, which has an average birth year of roughly 1980. Only U.S.-born individuals are displayed. Survey weights in both analytic samples are adjusted for selection into genotyping using inverse probability weighting.

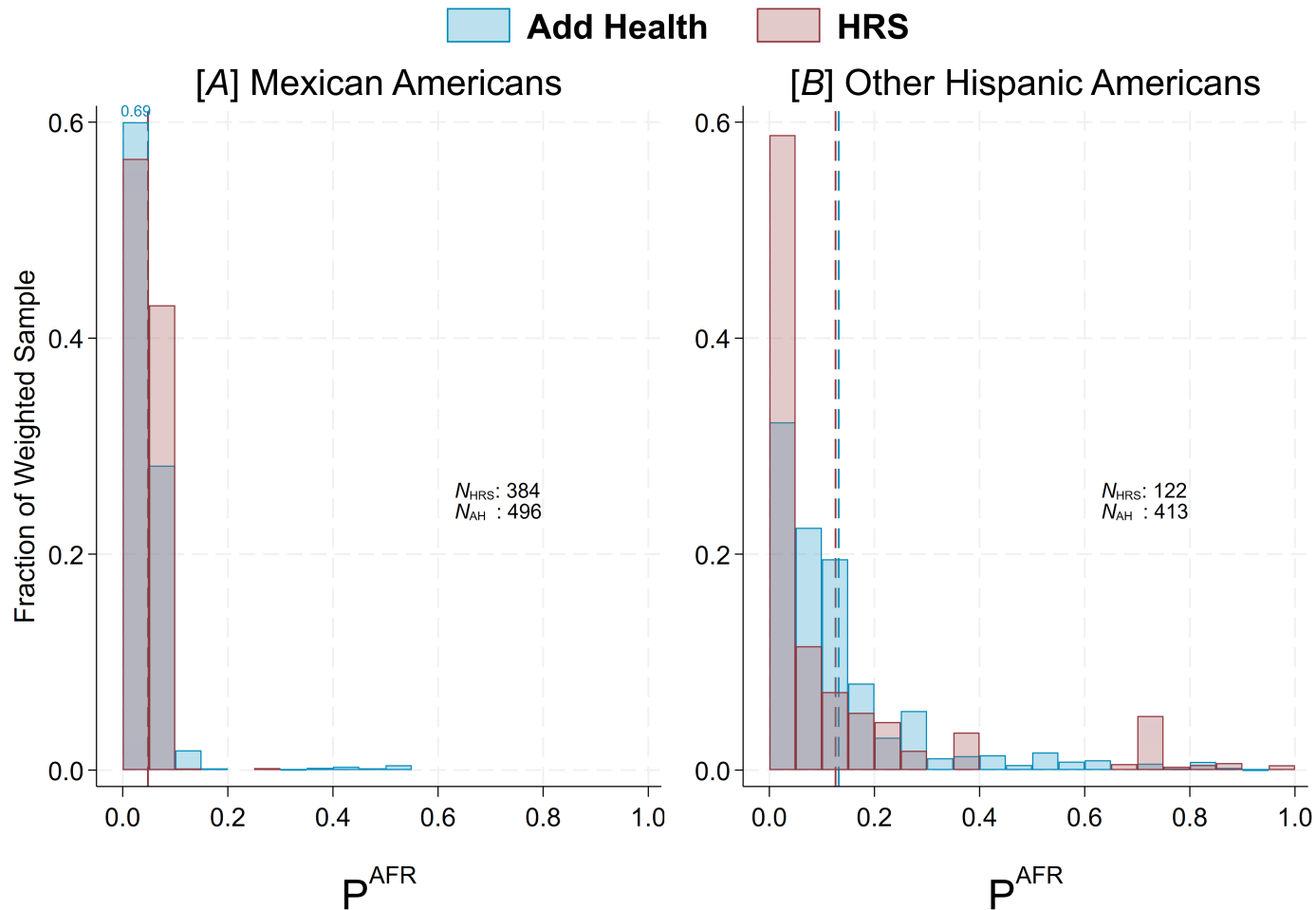


Fig. S3. Changes in Sub-Saharan African genetic similarity from 1945 to 1980 among Hispanic Americans.

This figure contains histograms of the distribution of the Sub-Saharan African genetic similarity proportion (P^{AFR}) separately for Mexican Americans and for Other Hispanic Americans. Genetic similarity proportions are estimated via supervised ADMIXTURE. The red bars display data from the Health and Retirement Study, which has an average birth year of approximately 1945; the blue bars display data from the Add Health Study, which has an average birth year of roughly 1980. Only U.S.-born individuals are displayed. Survey weights in both analytic samples are adjusted for selection into genotyping using inverse probability weighting.

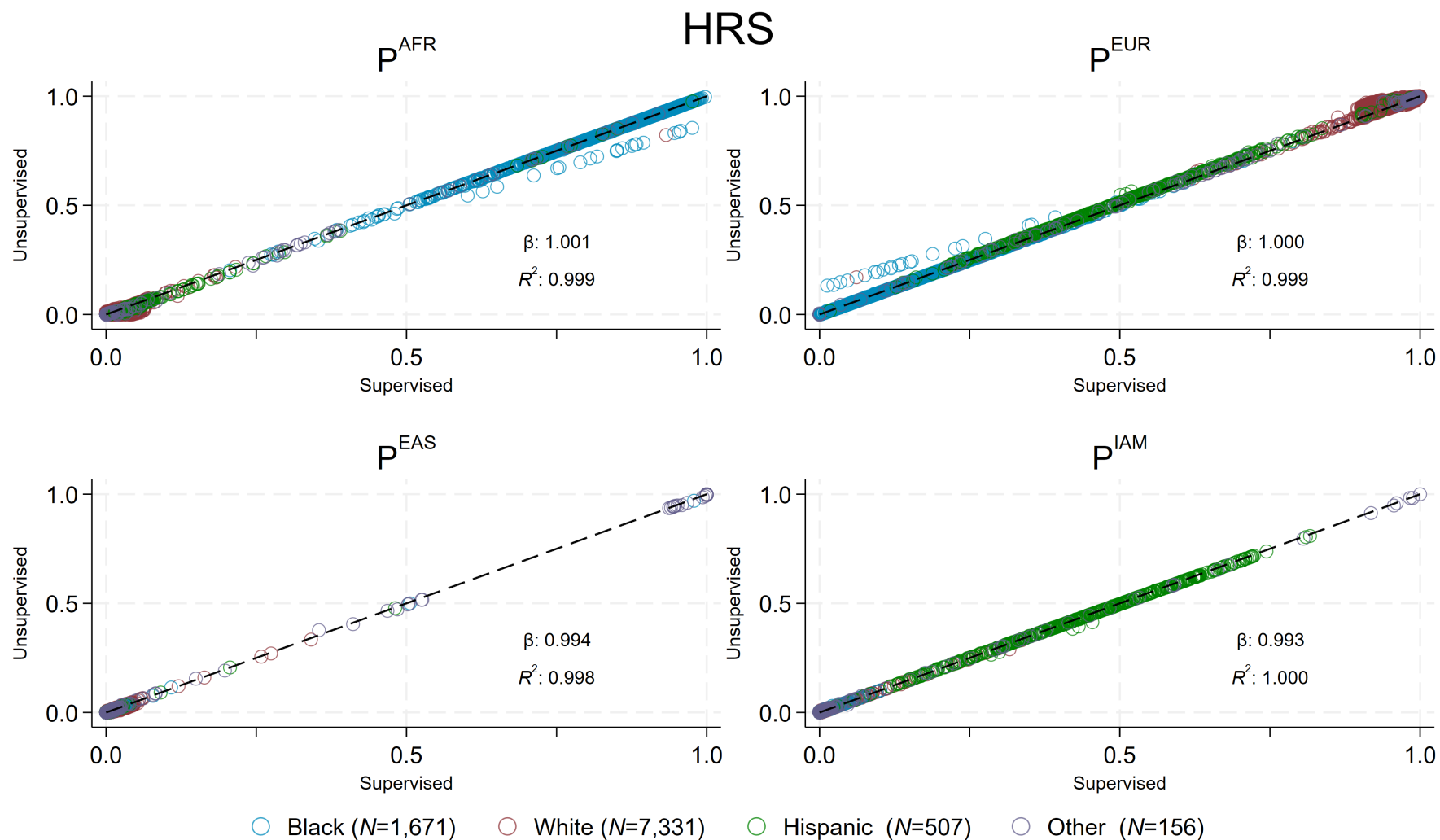


Fig. S4. Correspondence between supervised and unsupervised ADMIXTURE estimates in the Health and Retirement Study.

This figure uses scatter plots to compare the genetic similarity proportions of HRS respondents from supervised and unsupervised ADMIXTURE estimation ($K = 4$). Unsupervised ADMIXTURE estimates are derived using only the HRS sample. Markers are color-coded by respondent racial/ethnic identity.

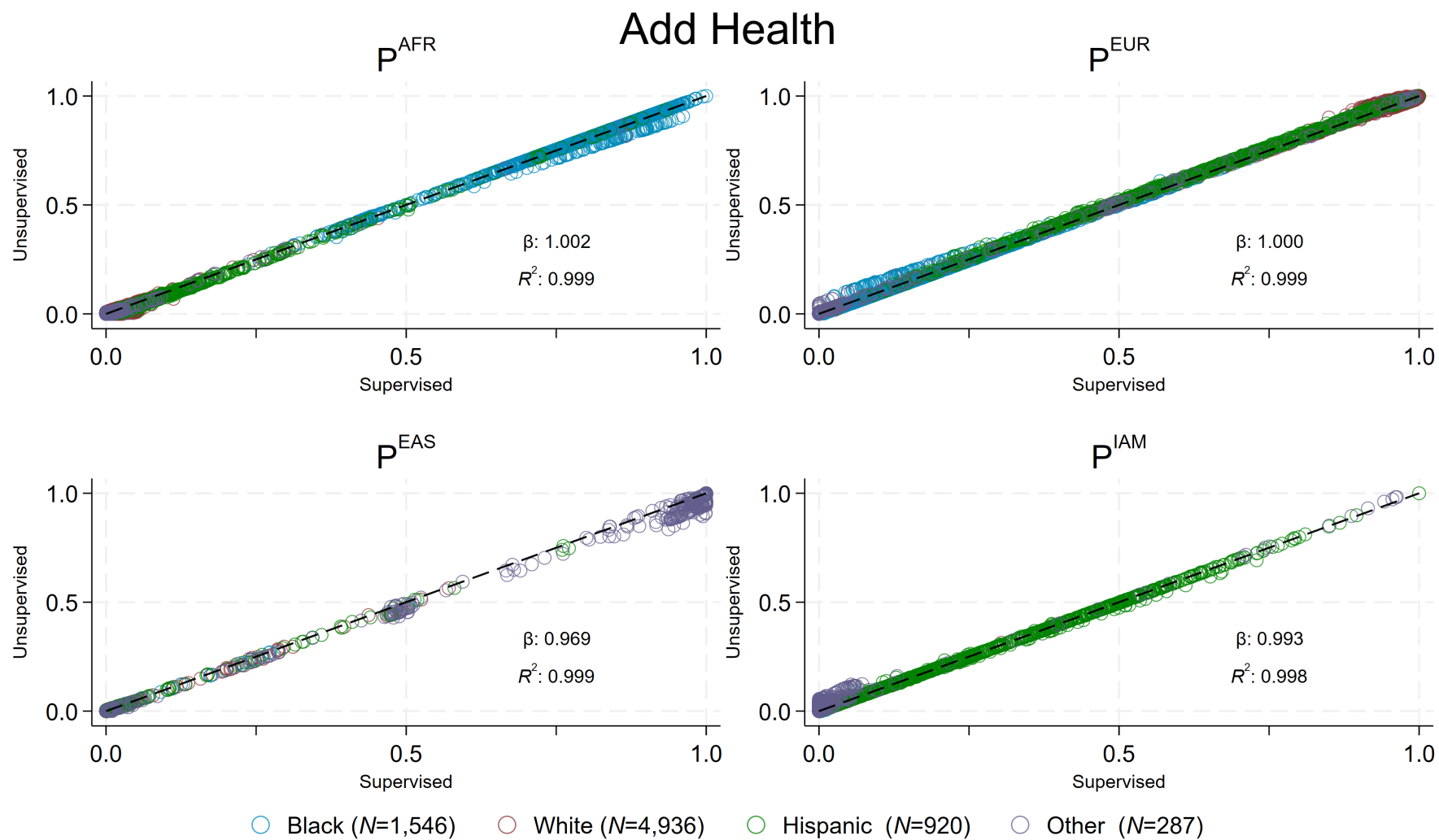


Fig. S5. Correspondence between supervised and unsupervised ADMIXTURE estimates in the Add Health study.

This figure uses scatter plots to compare the genetic similarity proportions of Add Health respondents from supervised and unsupervised ADMIXTURE estimation ($K = 4$). Unsupervised ADMIXTURE estimates are derived using only the Add Health sample. Markers are color-coded by respondent racial/ethnic identity.