

Genetic ancestry and social race are nearly interchangeable

Abstract

It has been claimed that social race and genetic ancestry are not closely related, or even unrelated. Here we test this claim by applying predictive modeling in both directions, i.e., predicting genetic ancestry from social race(s), and predicting social race(s) from genetic ancestry. We utilize the public Pediatric Imaging, Neurocognition, and Genetics (PING) dataset, so that others may examine the data as well.

In the simple scenario where we are only concerned with self-identified white, black, and mixed (black-white) race individuals (571 whites, 140 blacks, 25 mixed), model accuracy was very high. Predicting social race from genetic ancestry resulted in an area under curve (AUC) of .994, whereas predicting genetic ancestry from social race had a model R^2 adjusted of .992.

Using the full dataset, there are 8 census-type categories of social race. Using cross-validated multinomial regression to predict social race from 6 genetic ancestry variables, we find that the AUC is .89. Using Dirichlet regression to predict ancestries from social race, we find an overall correlation of .94 ($R^2 = 88.4\%$). Further analyses using more sophisticated methods found similar results. In conclusion, social race and genetic ancestry are nearly interchangeable.

Keywords: race, genetics, ethnicity

Introduction

There is no lack of books and articles arguing that race is a social construct (Evans, 2019; Gould, 1981; Montagu, 1942; Sussman, 2014). Representative headlines in the media include “Race Is Real, But It’s Not Genetic” from Discovery Magazine (Goodman, 2020), while in The Atlantic, we learn that “people’s racial identity may be statistically correlated with their ancestry, albeit unreliably” (Holmes, 2018), and in Scientific American that “Racial categories are weak proxies for genetic diversity and need to be phased out” (Gannon, 2016). There was an entire special issue in National Geographic about the supposed unreality or social construction of race (National Geographic, 2018; Nyborg, 2019). However, such works do not actually examine the strength of the statistical associations between social race and genetic ancestry or quote any statistics in the case of newspaper or magazine articles.¹ Thus, there is a need to quantify how strong the relationship is between social race and genetic ancestry.

Data

We used data from the Pediatric Imaging, Neurocognition, and Genetics (PING) dataset (<http://pingstudy.ucsd.edu/>) (Jernigan et al., 2016). This choice was motivated by the

¹ By *social race*, we mean here human-designated racial classification of persons, whether by themselves or by others.

availability of the dataset for public use. Though the dataset does not appear to be in the public domain or deliberately designated for public use, Noble et al (Noble et al., 2015) used the dataset in a study. As part of their publication, they attached large parts of the dataset to the journal website, thus making it freely available for others' use. The fact that the dataset is thus de facto public means that others will be able to replicate our analyses to verify they are correct or carry out follow-up analyses.

The dataset itself consists of 1,493 American children and youths (ages 3-20, mean 11.7) who underwent detailed phenotyping including surveys, neuroscientific (MRI), cognitive testing (NIH Toolbox Cognition Battery), and genetic testing. The subjects were recruited through "local postings and outreach activities conducted in the greater metropolitan areas of Baltimore, Boston, Honolulu, Los Angeles, New Haven, New York, Sacramento, and San Diego", and as such, are not perfectly representative of the American population of this age group. Only 1391 subjects were available in the public dataset.

As part of the interviewing, the child or their primary guardian was asked which of the following racial categories they identified with: 1) Hispanic or Latino, 2) Pacific Islander, Samoan, or Hawaiian, 3) Asian, 4) African American or Black, 5) American Indian or Native American, 6) White, or 7) Other. Thus, for every person, there is a set of 7 social race binary variables available for study. We coded the data two different ways: First, the standard simplified census-approach. In this approach, anyone who responds yes to Hispanic is classified as Hispanic. Anyone else who only selects a single option is classified as that. Anyone who selected multiple options was classified as multiracial. This produced 8 categories (the 7 available options + multiracial). Second, the common combinations with lumping. Every combination of chosen races is combined into a single compound group. Then all groups that were fewer than 20 subjects were lumped together in a remainder category. This approach resulted in 11 categories, shown in Table 1.

Group	Count	Percent
White	571	41.05
Remainder	177	12.72
African American	140	10.06
Hispanic, White	140	10.06
Asian	122	8.77
Hispanic	71	5.10
Asian, White	60	4.31
Pacific Islander, Asian, White	32	2.30
Hispanic, African American	29	2.08
African American, White	25	1.80
Other	24	1.73

Table 1. Distribution of social races by the common combinations coding with $n = 20$ as the minimum group size. Encoding was done by `forcats::fct_lump_min()`.

The genetic testing consisted of a standard microarray measurement (Illumina Human660W-Quad BeadChip, 550k variants). The PING group carried out ancestry analysis and assigned each subject's global ancestry to one of 6 large clusters: African, Central Asian, East Asian, European, Native American, and Oceanic. The estimation was done using ADMIXTURE. These variables were also released as part of the dataset by (Noble et al., 2015).

Results

Results are presented in two parts. In our first analysis, we examine only the African-European ancestry subset of the data. Then in the second part, we extend the analysis to the entire dataset.

African-Europeans

The first subset consists of subjects who selected only African or European races, and whose genetic ancestry of these two clusters sum to at least 95% ($n = 649$, 497 whites, 130 blacks, 22 mixed). For this dataset, the African and European ancestry components are nearly perfectly negatively correlated ($r = -1.00$), and thus the genetic data is effectively one-dimensional. The outcome variable is the ordered factor of social race with mixed being the intermediate level. Thus, in this simplified scenario, the data can be easily modeled. When predicting social race, ordinal logistic regression can be used, and in the case of predicting genetic ancestry, a simple linear regression is sufficient. The logistic model had an area under the curve (AUC) of .994 and a pseudo- R^2 of .932. Figure 1 shows the model results.

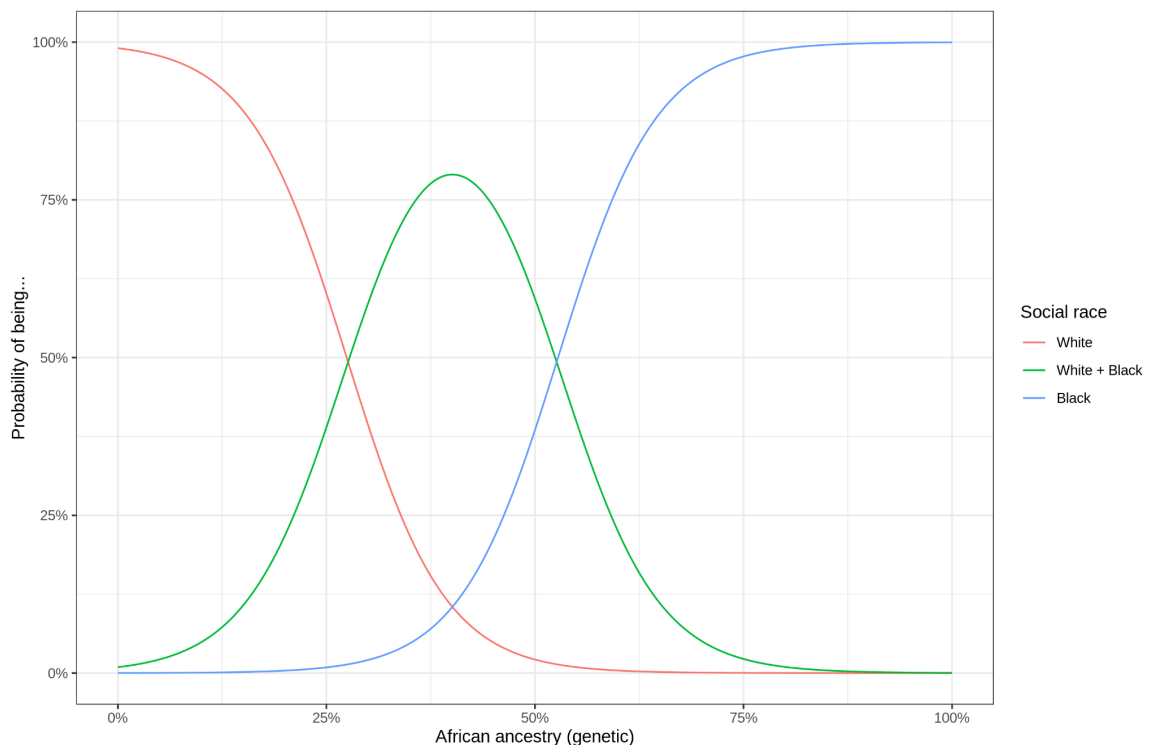


Figure 1. Probability of social race categories as a function of genetic ancestry.

In the figure, we see that the mixed race group is not entirely centered at 50%, as one might naively expect. The value is instead at 40% African and 60% European. This is probably best interpreted as the remnants of the one-drop rule, wherein any amount of African ancestry would classify a person as African by some US state laws. This replicates the prior results by (Lasker et al., 2019) which found the mixed black-white group had a mean European ancestry of 79.6%.

Conversely, predicting genetic ancestry from social race results in a well-fitting model with an adjusted R^2 of 0.955. This model is essentially just the average ancestry for each of the three groupings, as shown in Figure 2.

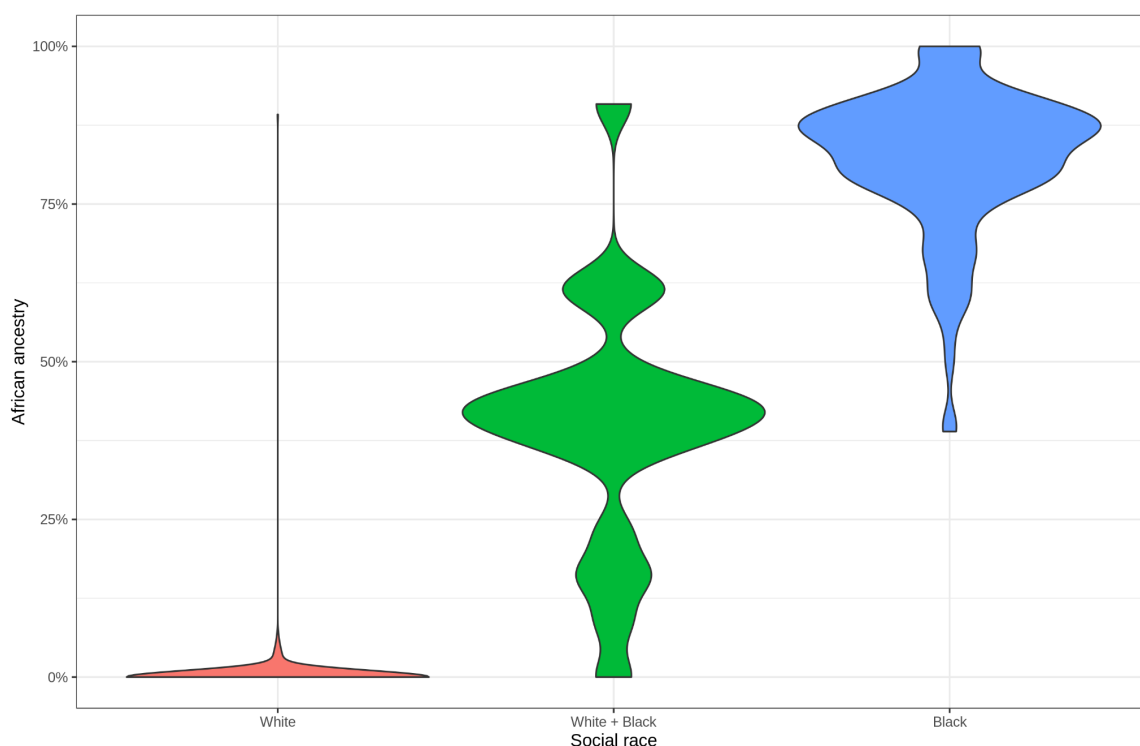


Figure 2. Violin plots of African ancestry by social race. The average ancestry proportions are .00, .40, and .82, respectively, for Whites, White+Blacks, and Blacks.

The full sample

With the results in mind from the previous section, we are now ready to examine the full dataset. There are now 11 categories to predict, and they cannot be coded as an ordinal variable. Thus, one cannot use logistic or ordinal regression. Multinomial regression is the standard approach for this kind of data. In this approach, the probability of a case belonging to each category is estimated based on the input variables, which in this case are the 6 genetic ancestry variables. We used the **nnet** implementation of this model as implemented in **tidymodels** (Kuhn et al., 2020; Ripley & Venables, 2021). To avoid overfitting, we used 20-fold cross validation. The estimated model accuracy was $AUC = .925$, with a strict concordance of 76.7% versus 41.0% by guessing the largest group.²

² We used the Hand and Till variant AUC generalization for multiclass data, as this was the default in **tidymodels**. https://yardstick.tidymodels.org/reference/roc_auc.html

The 6 genetic ancestries sum to 1, and thus using multivariate multiple regression is probably inappropriate because the predicted values are not constrained to [0, 1] (the same reason it is often inappropriate for binary outcomes). The standard approach to this is to use dirichlet regression, which is made to model such proportional data and accomplishes this using data rescaling (Douma & Weedon, 2019). Dirichlet regression is implemented in the DirichletReg package (Maier, 2021), which we used here to fit the data. Dirichlet regression does not provide any overall model fit, but one can examine the predicted values compared to the true values for each dimension, as shown in Figure 3.

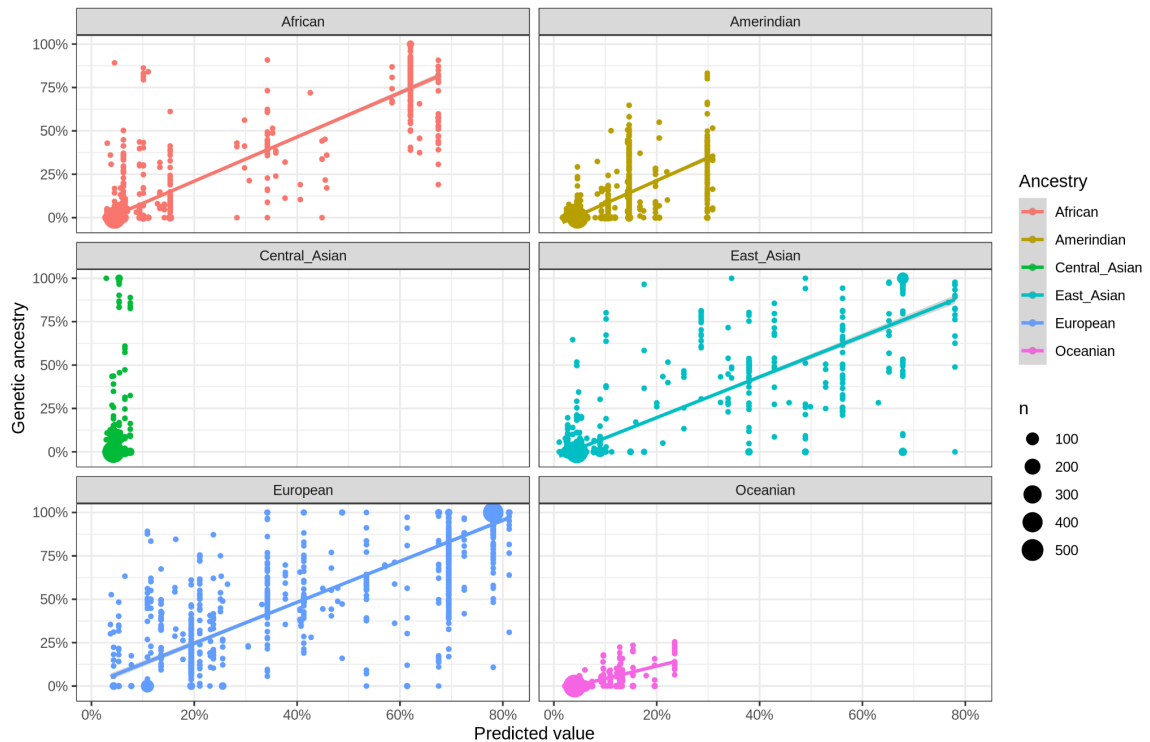


Figure 3. Model predictions from dirichlet regression for predicting 6 genetic ancestries.

The correlations for each ancestry are: African .92, Amerindian .74, Central Asian .19, East Asian .86, European .88, and Oceanian .77. Overall, the correlation between any prediction and the true value is .92. All values are far beyond chance levels. Few persons had Oceanian or Central Asian ancestry in our dataset, so it is not surprising the correlations are weaker for them. In addition, no person in our sample was an unadmixed Amerindian.

We tried some variations on the analyses in this section. First, we tried using the census simplified social race encodings instead, thereby reducing the group count to 8 and allowing for very small groups (there were only 4 American Indians). For predicting social race, the AUC was .915 and the concordance was 84.4% versus 41.0% expected by guessing the modal value. For predicting ancestry, the overall correlation was .90. These results are practically identical to the ones with the more complicated coding. Thus, the specific coding was not important for the strength of the results.

Second, it is possible that interactions between variables were important. To try to include these, we used a random forest model to predict social race as encoded by the common combinations. This model produced a model fit of AUC = 0.930, and concordance of 77.4% (versus 41% by guessing the modal value). Multivariate random forest to predict genetic ancestries from the social race variables produced an overall correlation of $r = .94$. Both results are very slightly better than those using the simpler additive models. Thus, we find that interactions are unimportant.

Third, to assess whether standard multiple regression produced inappropriate results, we fit the multivariate ancestry model. The results were mostly sensible, though some out of bounds predictions were produced (15% of values were below 0, none above 1). Overall, the model accuracy was more or less the same dirichlet regression, $r = .92$.

Discussion

We examined the statistical relationship between social race and genetic ancestry in a moderate-sized but diverse American sample of children and youths. Despite popular claims to the contrary, we found that the associations between the variables were extremely strong. When predicting social race, AUCs were consistently above .90. According to a guideline for the interpretation of AUC values from a statistics textbook, values above .90 are considered “outstanding” compared to merely “excellent” in the span .80 to .90 (Hosmer & Lemeshow, 2000, p. 162). In the simplified case where we could compute a pseudo-R², this was .932, again extremely strong. Such values are rarely encountered in applied research (Gignac & Szodorai, 2016; Nuijten et al., 2020). Whether or not we focused on the simplified situation of only African-European mixes, or whether we looked at the full sample, the accuracies remained very high. The empirical results stand in stark contrast to the various claims of weak or even nonexistent associations that we quoted earlier. The results in the present study were almost identical to those reported in two prior studies that carried out a similar, but more limited analysis (Fang et al., 2019; Lasker et al., 2019). Thus, it is not likely that our sample is an outlier among other samples.

Acknowledgements and supplementary resources

We wish to thank the PING consortium for their dataset. Special thanks to (Noble et al., 2015) for releasing the dataset to the public. The project files can be found at <https://osf.io/qxvg8/>, and the data can be downloaded from the journal website of Noble et al’s study (<https://www.nature.com/articles/nn.3983>).

References

- Douma, J. C., & Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution*, 10(9), 1412–1430.
<https://doi.org/10.1111/2041-210X.13234>
- Evans, G. (2019). *Skin Deep: Journeys in the Divisive Science of Race*. Oneworld Publications.

- Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T. L., Huang, J., Vujkovic, M., Damrauer, S. M., Pyarajan, S., Gaziano, J. M., DuVall, S. L., O'Donnell, C. J., Cho, K., Chang, K.-M., Wilson, P. W. F., Tsao, P. S., Gaziano, J. M., Ramoni, R., Breeling, J., ... Tang, H. (2019). Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *The American Journal of Human Genetics*, *105*(4), 763–772. <https://doi.org/10.1016/j.ajhg.2019.08.012>
- Gannon, M. (2016, February 5). Race Is a Social Construct, Scientists Argue. *Scientific American*.
<https://www.scientificamerican.com/article/race-is-a-social-construct-scientists-argue/>
- Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78.
<https://doi.org/10.1016/j.paid.2016.06.069>
- Goodman, A. (2020, June 25). *Race Is Real, But It's Not Genetic*. Discover Magazine.
<https://www.discovermagazine.com/planet-earth/race-is-real-but-its-not-genetic>
- Gould, S. J. (1981). *The mismeasure of man* (1st ed). Norton.
- Holmes, I. (2018, April 25). What Happens When Geneticists Talk Sloppily About Race. *The Atlantic*.
<https://www.theatlantic.com/science/archive/2018/04/reich-genetics-racism/558818/>
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed). Wiley.
- Jernigan, T. L., Brown, T. T., Hagler, D. J., Akshoomoff, N., Bartsch, H., Newman, E., Thompson, W. K., Bloss, C. S., Murray, S. S., Schork, N., Kennedy, D. N., Kuperman, J. M., McCabe, C., Chung, Y., Libiger, O., Maddox, M., Casey, B. J., Chang, L., Ernst, T. M., ... Dale, A. M. (2016). The Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository. *NeuroImage*, *124*(0 0), 1149–1154. <https://doi.org/10.1016/j.neuroimage.2015.04.057>
- Kuhn, M., Wickham, H., & RStudio. (2020). *tidymodels: Easily Install and Load the*

“*Tidymodels*” Packages (0.1.0) [Computer software].

<https://CRAN.R-project.org/package=tidymodels>

Lasker, J., Pesta, B. J., Fuerst, J. G. R., & Kirkegaard, E. O. W. (2019). Global Ancestry and Cognitive Ability. *Psych*, 1(1), 431–459.

<https://doi.org/10.3390/psych1010034>

Maier, M. J. (2021). *DirichletReg: Dirichlet Regression* (0.7-1) [Computer software].

<https://CRAN.R-project.org/package=DirichletReg>

Montagu, A. (1942). *Man’s Most Dangerous Myth: The Fallacy of Race* (1st ed.).

Columbia University Press.

National Geographic. (2018). *Black and White*.

<https://www.nationalgeographic.com/magazine/article/race-twins-black-white-biggs>

Noble, K. G., Houston, S. M., Brito, N. H., Bartsch, H., Kan, E., Kuperman, J. M., Akshoomoff, N., Amaral, D. G., Bloss, C. S., Libiger, O., Schork, N. J., Murray, S. S., Casey, B. J., Chang, L., Ernst, T. M., Frazier, J. A., Gruen, J. R., Kennedy, D. N., Van Zijl, P., ... Sowell, E. R. (2015). Family income, parental education and brain structure in children and adolescents. *Nature Neuroscience*, 18(5), 773–778. <https://doi.org/10.1038/nn.3983>

Nuijten, M. B., van Assen, M. A. L. M., Augusteijn, H. E. M., Cromptvoets, E. A. V., & Wicherts, J. M. (2020). Effect Sizes, Power, and Biases in Intelligence Research: A Meta-Meta-Analysis. *Journal of Intelligence*, 8(4), 36.

<https://doi.org/10.3390/jintelligence8040036>

Nyborg, H. (2019). Race as Social Construct. *Psych*, 1(1), 139–165.

<https://doi.org/10.3390/psych1010011>

Ripley, B., & Venables, W. (2021). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models* (7.3-16) [Computer software].

<https://CRAN.R-project.org/package=nnet>

Sussman, R. W. (2014). *The myth of race: The troubling persistence of an unscientific*

idea.