Submitted: 15th of November 2021
Published: 22nd of December 2021
DOI: 10.26775/OP.2021.12.22
ISSN: 2597-324X

Genetic ancestry and social race are nearly interchangeable

Emil O. W. Kirkegaard *

Abstract

It has been claimed that social race and genetic ancestry are at best weakly related. 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 (n = 1,391), 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, an overall accuracy (concordance) of 98.0 %, and a pseudo-R^2 of .951. Conversely, 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 = .884). Further analyses using more sophisticated methods (random forest, support vector machine) found similar results. In conclusion, social race and genetic ancestry are nearly interchangeable.

Keywords: race, genetics, ethnicity

1 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 recently 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 or report the strength of the statistical associations between social race and genetic ancestry.[1] Similarly, surveys of geneticists and anthropologists do not find any consensus about how strong the relationship is (Nelson et al., 2018; Wagner et al., 2017). Thus, there is a need to quantify how strong the relationship is between social race and genetic ancestry.

2 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 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. (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

*Ulster Institute for Social Research, United Kingdom, Email: emil@emilkirkegaard.dk

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

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 1,391 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
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.

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()*.

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

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 ancestries to 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).

3 Results

Results are presented in two parts. In the first part, we examine only the African-European ancestry subset of the data. Then in the second part, we extend the analysis to the entire dataset. All analyses were done with R 4.1.2.

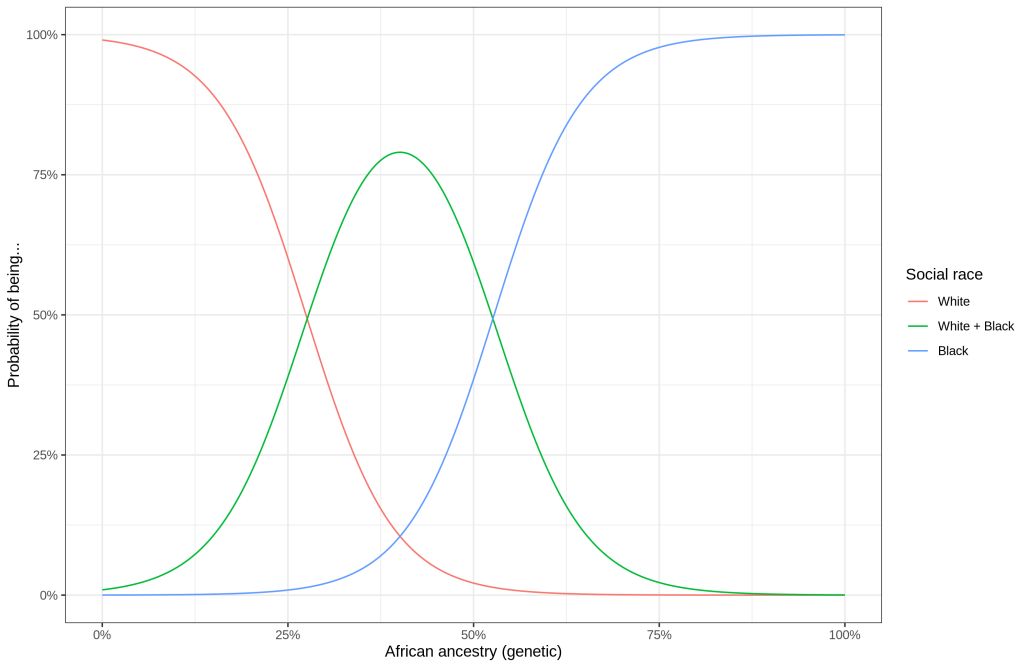

Figure 1: Probability of social race categories as a function of genetic ancestry.

Table 2: Confusion matrix for model predictions.

	Predicted values		
True values	White	White + Black	Black
White	496	0	1
White + Black	5	13	4
Black	0	3	127

3.1 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 the mixed group being the intermediate level. Thus, in this simplified scenario, the data can be modeled using ordinal logistic regression. For 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.

In the figure, we see that the mixed race group is not entirely centered at 50 %, as one might naively expect. The modal value is instead at 40 % African and 60 % European. We can think of two explanations for this. The first is that it is a remnant of the one-drop rule, wherein any amount of African ancestry would classify a person as African by some US state laws (Liz, 2018). The second is that many individuals identifying (or by their parents) are first generation mixes between an African American and a White. Since African Americans have about 80 % African ancestry, and White Americans about 0 %, the offspring will have about 40 %, which is the modal value observed (thanks for Gerhard Meisenberg for this suggestion). Either way, our finding replicated the prior results by (Lasker et al., 2019) which found the mixed Black-White group was intermediate in European ancestry, and had a mean European ancestry of 79.6 %.

With regards to the model predictions, it is informative to look at the confusion matrix, which shows the concordance between model predictions and true values. This is shown in Table 2.

Overall, the concordance was 98.0 %. It was 99.8 % for Whites, 59.1 % for mixed, and 97.7 % for Blacks. These results thus closely match those found by Tang et al. (2005). If we reduce the modeling outcome to just predicting whether a subject reported being Black or not, the concordance was 99.1 %, AUC = .995, and

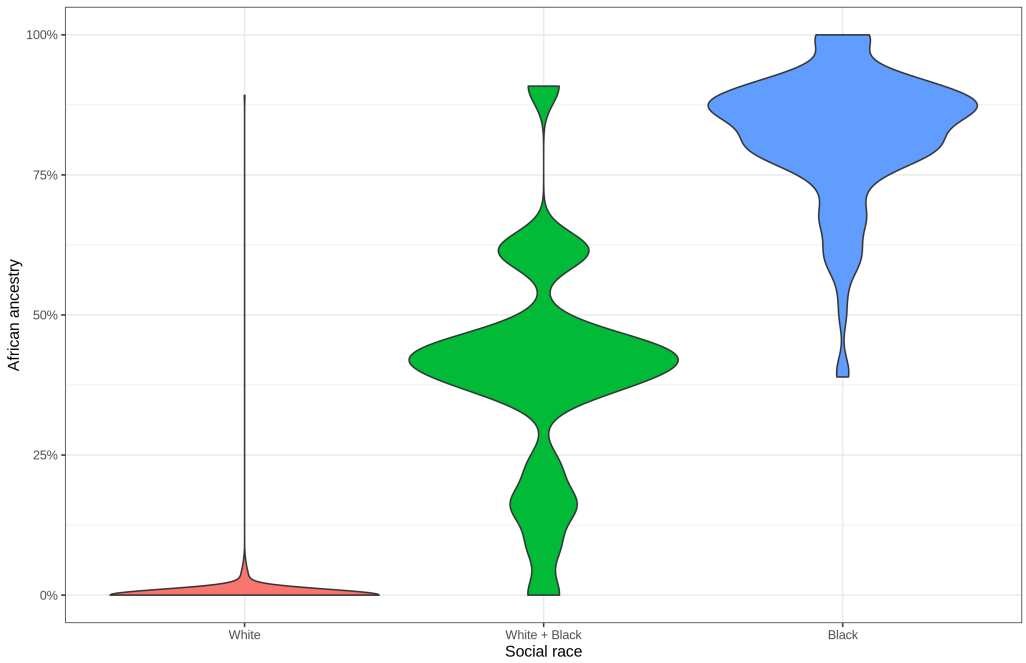

Figure 2: Violin plots of African ancestry by social race. The average ancestry proportions are .00, .40, and .82, respectively, for Whites, White + Blacks, and Blacks.

pseudo-R^2 .951. Thus, the main difficult of the model comes from telling the mixed from the Blacks, which can also be seen in the confusion matrix.

Conversely, predicting genetic ancestry from social race requires merely a linear model. The fit is excellent, with an adjusted R^2 of .955. This model is just the average ancestry for each of the three groupings, as shown in Figure 2.

3.2 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 parametric 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 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]

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], nor do they necessarily sum to 1. 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 akin to R^2, but one can examine the predicted values compared to the true values for each dimension, as shown in Figure 3.

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, which thus corresponds to an R^2 of .85. Few persons had Oceanian or Central Asian ancestry in our dataset, so it is not surprising the correlations are weaker for them, though all are far beyond chance levels (all p's $< 10^{-13}$). In addition, no person in our sample was an unadmixed Amerindian.

[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

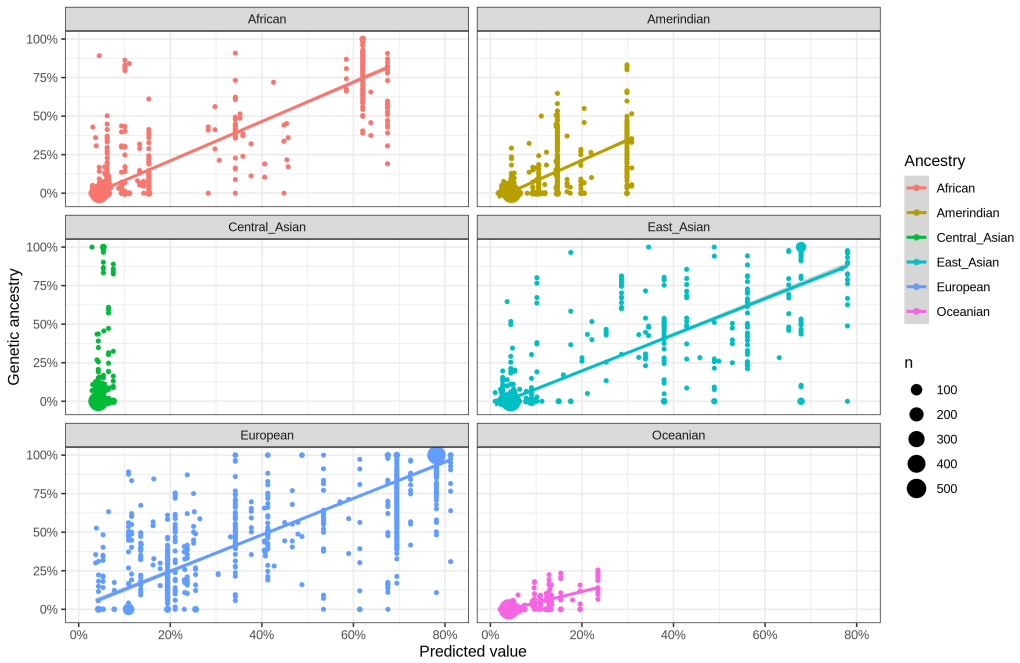

Figure 3: Model predictions from Dirichlet regression for predicting 6 genetic ancestries.

We tried some variations on the analyses in this section. First, we tried using the standard 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 nonlinear relationships or interactions between variables were important. To try to capture these effects, 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. Similarly, using a radial support vector machine to predict social race, we attained a concordance of 75.7 %, and AUC of .851. Thus in general, we don't find that nonlinear effects or interactions are important.

Third, to assess whether standard multiple regression produced inappropriate results, we fit the multivariate ordinary least squares 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.

4 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 scenario of only African-European mixes where we could compute a pseudo-R^2, this was .932, again extremely strong. Such values are rarely encountered in applied research with individuals (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 nonexistent associations that we quoted earlier. The results in the present study were very similar to those reported in prior studies that carried out a

similar, but more limited analysis (Fang et al., 2019; Lasker et al., 2019; Tang et al., 2005). Thus, it is not likely that our sample is an outlier among other samples.

5 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. (2015)'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. doi: 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., ... Tang, H. (2019). Harmonizing genetic ancestry and self-identified race/ethnicity in genome-wide association studies. *Am. J. Hum. Genet.*, *105*(4), 763–772. doi: 10.1016/j.ajhg.2019.08.012

Gannon, M. (2016, February 5). *Race is a social construct, scientists argue.* Retrieved from 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. *Pers. Individ. Dif.*, *102*, 74–78. doi: 10.1016/j.paid.2016.06.069

Goodman, A. (2020, June 25). Race is real, but it's not genetic. *Discover Magazine*. Retrieved from 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*. Retrieved from 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., Jr, Akshoomoff, N., Bartsch, H., Newman, E., ... Pediatric Imaging, Neurocognition and Genetics Study (2016). The pediatric imaging, neurocognition, and genetics (PING) data repository. *Neuroimage*, *124*(Pt B), 1149–1154. doi: 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*. Retrieved from 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. doi: 10.3390/psych1010034

Liz, J. (2018). "the fixity of whiteness": Genetic admixture and the legacy of the one-drop rule. *Critical Philosophy of Race*, *6*(2), 239–261.

Maier, M. J. (2021). Dirichletreg: Dirichlet regression (0.7-1). *Computer software*. Retrieved from 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.* Retrieved from https://www.nationalgeographic.com/magazine/article/race-twins-black-white-biggs

Nelson, S. C., Yu, J.-H., Wagner, J. K., Harrell, T. M., Royal, C. D., & Bamshad, M. J. (2018). A content analysis of the views of genetics professionals on race, ancestry, and genetics. *AJOB Empir. Bioeth.*, *9*(4), 222–234. doi: https://doi.org/10.1080/23294515.2018.1544177

Noble, K. G., Houston, S. M., Brito, N. H., Bartsch, H., Kan, E., Kuperman, J. M., ... Sowell, E. R. (2015). Family income, parental education and brain structure in children and adolescents. *Nature Neuroscience*, *18*(5), 773–778. doi: 10.1038/nn.3983

Nuijten, M. B., van Assen, M. A. L. M., Augusteijn, H. E. M., Crompvoets, 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. doi: 10.3390/jintelligence8040036

Nyborg, H. (2019). Race as social construct. *Psych*, *1*(1), 139–165. doi: 10.3390/psych1010011

Ripley, B., & Venables, W. (2021). nnet: Feed-forward neural networks and multinomial log-linear models (7.3-16). *Computer software*. Retrieved from https://CRAN.R-project.org/package=nnet

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

Tang, H., Quertermous, T., Rodriguez, B., Kardia, S. L. R., Zhu, X., Brown, A., ... Risch, N. J. (2005). Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am. J. Hum. Genet.*, *76*(2), 268–275. doi: 10.1086/427888

Wagner, J. K., Yu, J.-H., Ifekwunigwe, J. O., Harrell, T. M., Bamshad, M. J., & Royal, C. D. (2017). Anthropologists' views on race, ancestry, and genetics: WAGNER et al. *Am. J. Phys. Anthropol.*, *162*(2), 318–327. doi: 10.1002/ajpa.23120