Submitted: 26th of July 2016
Published: 12th of September 2016

Does sub-European genomic ancestry predict outcomes for US states?

Emil O. W. Kirkegaard*

Open Quantitative Sociology & Political Science

Abstract

Estimates of sub-European ancestry among European Americans by US state were obtained from a recent study of customers of the personal genomics company 23andme (N=148,789). The ancestry estimates were used to attempt to predict cognitive ability and socioeconomic performance across US states (N=50). However, results indicated that they had little or no reliable predictive validity, which was also the case when related ancestries were combined to increase precision.

Keywords: ancestry, admixture, European, White, United States, general socioeconomic factor, S- factor, NAEP, cognitive ability, intelligence, 23andme, LASSO regression

1 Introduction

In recent years economists have started examining the predictive power of ancestry for socioeconomic outcomes[1]. Putterman & Weil (2010) estimated "the share of the year 2000 population in every country that is descended from people in different source countries in the year 1500". Their analyses of this dataset revealed that present day GDP per capita was strongly predictable from the ancestries of persons living in the countries. Fulford et al. (2016) estimated the ancestry proportions (mostly sub-European) of US counties over time based on consecutive census data. They showed that ancestry proportions are predictive of socioeconomic outcomes such as GDP per capita and generalized trust. This study is important because the use of longitudinal data means that the design automatically controls for county-level fixed effects such as climate and geographical location (which has been advocated by e.g. León & Burga-León (2015)).

Differential psychologists have long been interested in the predictive power of ancestry (Galton, 1869; Shuey, 1966; Jensen, 1969; Rushton & Jensen, 2005; Lynn, 2006). Their focus has, however, been different. The research has centered on the power of ancestry to explain observable group differences in cognitive ability[2] especially that between European and African Americans. Shuey (1966) reviewed most if not all the early studies. These studies used a variety of proxies to classify African Americans into those with more or less African/European ancestry: skin brightness/darkness, nose width, hair type and blood type. Using such non-genomic measures as estimates of genomic ancestry is problematic and the samples sizes were often small, so not much can be concluded from this early evidence (Loehlin et al., 1975).

Following the availability of genomic methods to estimate ancestry (Shriver et al., 2003) (Shriver et al., 2003), studies reporting relationships between variables and genomic ancestry are now commonplace[3]. Fuerst & Kirkegaard (2016a) compiled estimates from such studies to estimate the genomic ancestry of every nation (N=35) in the Americas and sub-national regions of the United States, Mexico, Colombia and

* Ulster Institute for Social Research. E-mail: emil@emilkirkegaard.dk

1 Socioeconomic is here used in a catch-all meaning which includes most if not all valued sociological outcomes such as income, educational attainment, health, and (lower) crime. See Kirkegaard (2014b); Kirkegaard & Fuerst (2014).

2 In this paper I will use cognitive ability to refer to the kind of mental ability called forth by standardized intelligence tests and scholastic tests such as the NAEP employed in this paper. This usage is in line with previous research (Lynn & Vanhanen, 2012).

3 Kirkegaard et al. (In Review) meta-analyzed studies that examined the relationship between genomic ancestry and socioeconomic outcomes at the individual-level for persons living in the Americas. They found that European ancestry was robustly associated with better outcomes (random effects mean r = .17, K = 27, total N ≈ 35,200).

Brazil (total N=169). They showed that European ancestry is a strong predictor of countries' relative standing on measures of cognitive ability and socioeconomic performance (mean r's .708 and .643, respectively). These analyses grouped all European genomic ancestry into one cluster. However, it has been noted that there is substantial variation in the mean levels of cognitive ability and socioeconomic performance for European Americans across US states which is possibly related to more fine-grained ancestry differences between the states. The goal of this study is to examine whether genomic measures of sub-European ancestry of US states can predict outcomes.

2 Method

2.1 Ancestry data

Bryc et al. (2015) estimated the genomic ancestry of customers of the personal genomics company 23andme (http://23andme.com/). They published the estimated proportions of ancestry for African Americans, Hispanic Americans and European Americans for each state with sufficient data. The genomic estimates for African and Hispanic Americans has been examined in a recent study (Kirkegaard & Fuerst, 2016), but the estimates of sub-European ancestry has not been examined in the context of sociology or differential psychology so far to my knowledge. This is perhaps because the published paper did not present the data in tabular form, only in the form of maps which are shown in Figure 1. However, lead author Kasia Bryc was kind enough to provide the numerical data to me[4].

These maps may usefully be compared to the maps produced by Fulford et al. (2016) from census data which are shown in Figure 2.

It can be seen that there is broad agreement with regards to the distribution of Scandinavian/Nordic, Italian/Sardinian and German ancestries.

With regards to the representativeness of the 23andme-based data, it should be noted that this is a self-selected group (customers) of persons interested in personal genomics, and which is thus likely smarter and more well-off than the average person. This likely has the effect of skewing the ancestries towards the higher ability and more well-off groups.

2.2 Outcome data

Cognitive ability was measured using the NAEP (National Assessment of Educational Progress) for self-identified Whites as done in Kirkegaard & Fuerst (2016). The scores were copied from that paper.

General socioeconomic performance (S factor) for self-identified Whites was measured as the common factor of the three indicators provided by Measure of America (http://measureofamerica.org/): proportion of population with at least a bachelor's degree, average life-span and median income. The scores were copied from Kirkegaard & Fuerst (2016).

[4] Personal e-mail dated February 2016. The data are now also available on her website http://kasiabryc.com/datasets/.

Table 1: Correlations between ancestries and outcomes. Weighted by population. Numbers in brackets are 95 % analytic confidence intervals. The weighted correlation between CA and S is .70.

Ancestry	CA	S
Ashkenazi	0.31 [0.04 0.54]	0.55 [0.31 0.72]
Balkan	0.39 [0.12 0.60]	0.33 [0.06 0.56]
British/Irish	-0.41 [-0.62 -0.14]	-0.34 [-0.57 -0.07]
East European	0.40 [0.14 0.61]	0.18 [-0.11 0.43]
Finnish	0.11 [-0.17 0.38]	0.11 [-0.17 0.38]
French/German	-0.14 [-0.40 0.15]	-0.47 [-0.66 -0.22]
Iberian	0.06 [-0.22 0.33]	0.35 [0.08 0.57]
Italian	0.39 [0.13 0.60]	0.44 [0.18 0.64]
Middle Eastern	0.28 [0.00 0.52]	0.55 [0.32 0.72]
Sardinian	0.04 [-0.24 0.32]	-0.11 [-0.38 0.18]
Scandinavian	0.06 [-0.22 0.33]	-0.02 [-0.30 0.26]

3 Analyses

The data were analyzed using weighted (population size) multiple correlation and regression as done in the previous studies (Fuerst & Kirkegaard, 2016a,b; Kirkegaard & Fuerst, 2016).

3.1 Correlations

Table 1 gives the correlations between the ancestry variables and the two outcome variables. Examining the correlations, we see some with substantial values and confidence intervals that do not overlap zero, but many others are small or unexpected. For instance, British/Irish is strongly negative, Scandinavian is about 0, and Balkan moderately positive.

3.2 OLS regression

OLS regression provides a single best guess estimate for every predictor, in our case, each ancestry. The downside is that it tends to overfit models, especially those with many predictors. Tables 2 and 3 show the regression results for cognitive ability and S, respectively.

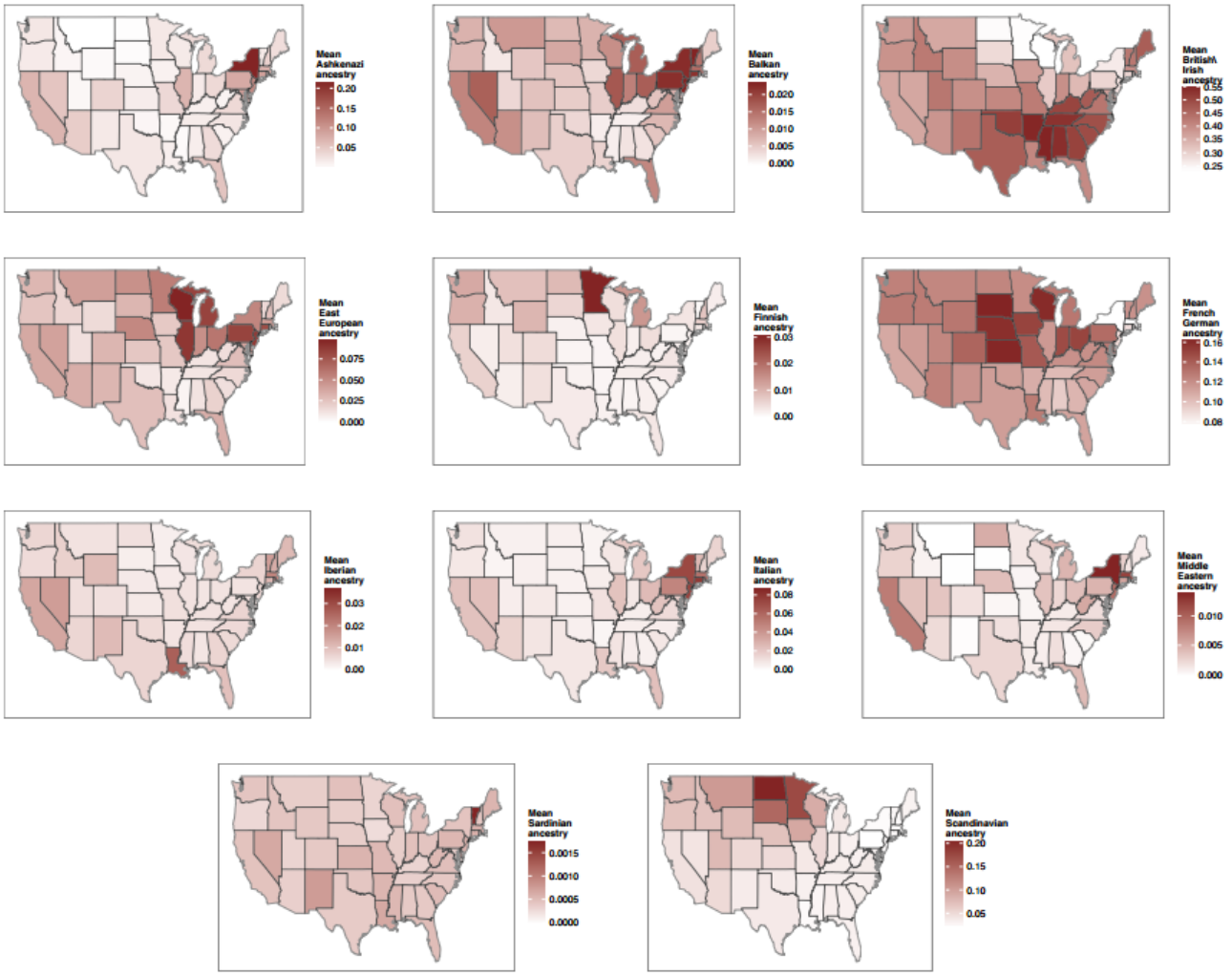

Figure 1: Sub-European ancestry among self-identified European Americans. From Figure S10 in Bryc et al. (2015).

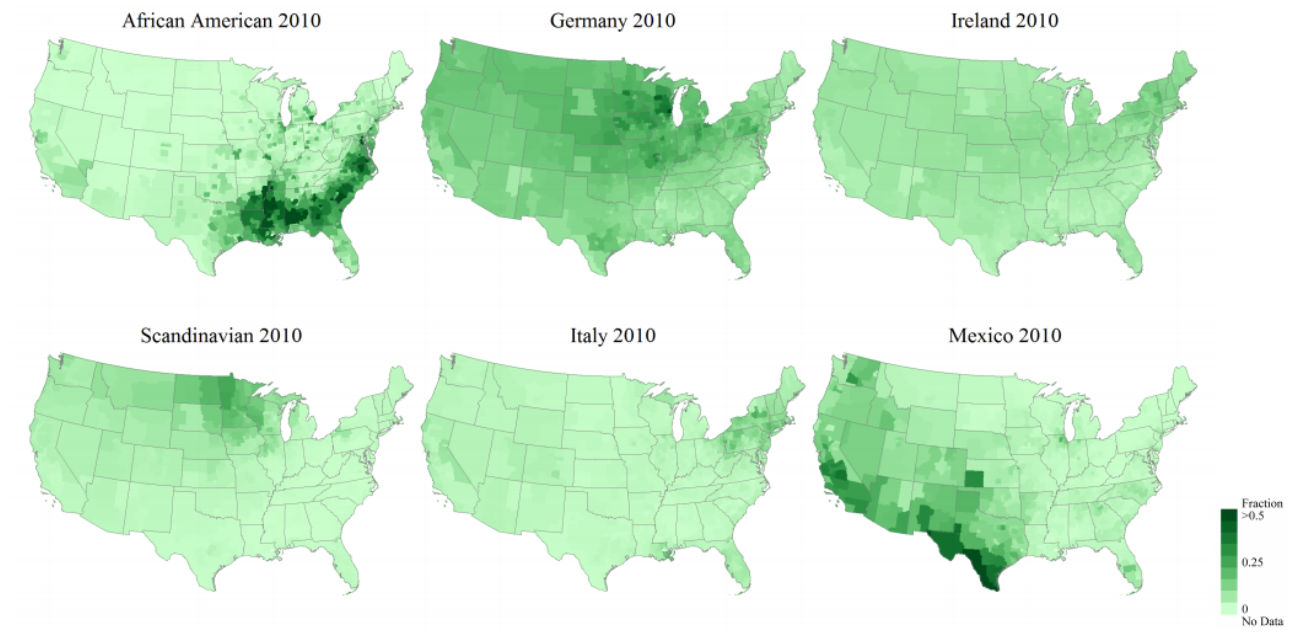

Figure 2: Estimated ancestry among the total population based on consecutive census data. From Fulford et al. (2016).

Table 2: OLS weighted regression standardized betas for cognitive ability among European American US states. N=50. R^2 = .31. Cross-validated R^2 = -3.16.[5] CI = 95 % analytic confidence intervals.

Predictor	Beta	SE	CI lower	CI upper
Ashkenazi	1.02	0.65	-0.29	2.34
Balkan	0.29	0.41	-0.53	1.12
British/Irish	2.70	1.56	-0.46	5.85
East European	1.13	0.60	-0.08	2.33
Finnish	0.18	0.24	-0.31	0.67
French/German	1.05	0.68	-0.32	2.43
Iberian	0.75	0.42	-0.10	1.60
Italian	1.13	0.53	0.06	2.21
Middle Eastern	-0.16	0.28	-0.73	0.40
Sardinian	-0.14	0.21	-0.56	0.28
Scandinavian	1.79	0.92	-0.06	3.64

Table 3: OLS weighted regression standardized betas for S factor among European American US states. N=50. R^2 = .48. Cross-validated R^2 = -1.06. CI = 95 % analytic confidence intervals.

Predictor	Beta	SE	CI lower	CI upper
Ashkenazi	0.54	0.58	-0.64	1.72
Balkan	-0.08	0.37	-0.82	0.66
British/Irish	0.28	1.40	-2.56	3.11
East European	0.31	0.54	-0.77	1.39
Finnish	-0.01	0.22	-0.45	0.44
French/German	-0.08	0.61	-1.31	1.16
Iberian	0.57	0.38	-0.20	1.33
Italian	-0.14	0.48	-1.11	0.83
Middle Eastern	0.04	0.25	-0.46	0.55
Sardinian	-0.14	0.19	-0.52	0.23
Scandinavian	0.40	0.82	-1.27	2.06

Both models were woefully overfit as demonstrated by the negative cross-validated R^2s. This indicates that other methods must be used to better estimate the predictive value of the predictors, if any.

3.3 *LASSO regression*

LASSO regression is a more conservative method, but has the benefit that it tends to provide better estimates of true predictive validity than OLS regression (James et al., 2013). Because LASSO regression has a stochastic element[6], it was run 500 times for each model.

The results indicated that no predictor was consistently useful in predicting the outcomes. Only for S were two predictors even remotely close to being reliable predictors, namely Ashkenazi and Middle Eastern ancestry which both had positive betas in about 37 % of the runs.

3.4 *Broader ancestries*

In reviewing the paper, Noah Carl suggested combining some of the ancestries to increase the power/precision (Cumming, 2012)[7]. This was done by combining Italian and Iberian to Southern, and combining Scandinavian and Finnish to Nordic. Both OLS and LASSO regression was then rerun. The results, however, were very similar: the models were vastly overfit to the data (negative cross-validated R^2's) and LASSO did not find any robust predictors.

4 Discussion and conclusion

The present study failed to find convincing evidence for predictive validity of sub-European ancestry among US states. Such validity is expected on theoretical grounds. If different sub-European ancestries differ in mean cognitive ability (Lynn & Vanhanen, 2012) or other important traits that cause variation in socioeconomic outcomes (Gottfredson, 1997; Murray, 2002; Rowe et al., 1998; Strenze, 2007), and migrants retain their cognitive ability and other traits when they relocate (spatial transferability hypothesis, see e.g. (Kirkegaard, 2014a)), then one would expect to find that the relative ancestry composition of states predicts their cognitive ability and S levels, all else equal. The problem is of course that all else is not equal. For instance, there has been substantial cognitive ability related migration within sub-European clusters in the US for many decades, and both emigration and immigration is selective to some degree which also possibly varies by country or region of origin. For these reasons, the relationships between ancestries and outcomes are theoretically expected to be quite noisy.

Aside from the above problems, the study is primarily limited by two facts. First, the self-selected nature of the sample limits its representativeness. However, unless this selection was different for each state, it should not in general throw off the results because the analyses do not depend on the mean levels of ancestry. To see this, imagine that Ashkenazi ancestry is associated with higher cognitive ability and better socioeconomic outcomes (Cochran et al., 2006; Lynn, 2011). If the 23andme estimates are based on persons

[5] The cross-validated R^2 were calculated using 10-fold cross validation (James et al., 2013).

[6] This is due to the use of cross-validation to choose the value of the shrinkage parameter.

[7] The standard error of each predictor is a function of the sample size and the intercorrelations with the other predictors. By this reducing the number of predictors, the standard error should decrease and hence the power increase.

with higher cognitive ability/S than the general population (positive selection), then one would expect that the mean level of Ashkenazi ancestry would be somewhat higher than among the general population. Still, the relative differences between states would be roughly the same which is sufficient for the employed methods to work.

Second, the present dataset only has a sample size of 50. If the relationships between sub-European ancestry and outcomes are weak to moderate, then the sample size may simply be too small to detect them. Given that effects of ancestry were found using non-genomic ancestry estimates, which are less precise, in the US using a much larger dataset (N=1154, (Fulford et al., 2016, see footnote 1), this seems plausible.

Supplementary material and acknowledgments

Data files and analysis code are available at the Open Science Framework repository https://osf.io/3nghx/.

The peer review thread can be found on the journal forum http://openpsych.net/forum/showthread.php?tid=288&pid=4161.

Thanks to Noah Carl, Bryan Pesta and Gerhard Meisenberg for reviewing the paper.

Thanks to Kasia Bryc for supplying the admixture data in numeric form.

References

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., & Mountain, J. L. (2015). The genetic ancestry of african americans, latinos, and european americans across the united states. *The American Journal of Human Genetics*, *96*(1), 37–53. doi: http://doi.org/10.1016/j.ajhg.2014.11.010

Cochran, G., Hardy, J., & Harpending, H. (2006). Natural history of ashkenazi intelligence. *Journal of biosocial science*, *38*(05), 659–693. doi: http://doi.org/10.1017/S0021932005027069

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Fuerst, J., & Kirkegaard, E. O. W. (2016a). Admixture in the americas: Regional and national differences. *Mankind Quarterly*, *56*, 255–373.

Fuerst, J., & Kirkegaard, E. O. W. (2016b). The genealogy of differences in the americas. *Mankind Quarterly*, *56*(3), 425–481.

Fulford, S. L., Petkov, I., & Schiantarelli, F. (2016). Does it matter where you came from? ancestry composition and economic performance of us counties, 1850-2010. *Ancestry Composition and Economic Performance of US Counties*. (Retrieved from http://papers.ssrn.com/abstract=2608567)

Galton, F. (1869). *Hereditary genius*. Macmillan and Company. (Retrieved from http://www.galton.org/books/hereditary-genius/index.html)

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*(1), 79–132. doi: http://doi.org/10.1016/S0160-2896(97)90014-3

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R* (Vol. 6). Springer.

Jensen, A. (1969). How much can we boost IQ and scholastic achievement. *Harvard educational review*, *39*(1), 1–123.

Kirkegaard, E. O. W. (2014a). Crime, income, educational attainment and employment among immigrant groups in norway and finland. *Open Differential Psychology*.

Kirkegaard, E. O. W. (2014b). The international general socioeconomic factor: Factor analyzing international rankings. *Open Differential Psychology*.

Kirkegaard, E. O. W., & Fuerst, J. (2014). Educational attainment, income, use of social benefits, crime rate and the general socioeconomic factor among 70 immigrant groups in denmark. *Open Differential Psychology*. (Retrieved from http://openpsych.net/ODP/2014/05/educational-attainment-income-use-of-social-benefits-crime-rate-and-the-general-socioeconomic-factor-among-71-immmigrant-groups-in-denmark/)

Kirkegaard, E. O. W., & Fuerst, J. (2016). Socioeconomic inequality in the US: Ethnicity, racial admixture and environmental causes. *Mankind Quarterly*.

Kirkegaard, E. O. W., Wang, M., & Fuerst, J. (In Review). Biogeographic ancestry and socioeconomic outcomes in the americas: a meta-analysis. *bioRxiv*, 055681.

León, F. R., & Burga-León, A. (2015). How geography influences complex cognitive ability. *Intelligence*, *50*, 221–227. doi: http://doi.org/10.1016/j.intell.2015.04.011

Loehlin, J. C., Lindzey, G., & Spuhler, J. N. (1975). *Race differences in intelligence*. W.H.Freeman & Co Ltd.

Lynn, R. (2006). *Race differences in intelligence: An evolutionary analysis.* Washington Summit Publishers.

Lynn, R. (2011). *The chosen people: A study of jewish intelligence and achievement*. Washington Summit Publishers.

Lynn, R., & Vanhanen, T. (2012). *Intelligence: A unifying construct for the social sciences* (1st ed.). Ulster Institute for Social Research.

Murray, C. (2002). IQ and income inequality in a sample of sibling pairs from advantaged family backgrounds. *The American Economic Review*, *92*(2), 339–343.

Putterman, L., & Weil, D. N. (2010). Post-1500 population flows and the long run determinants of economic growth and inequality. *The Quarterly Journal of Economics*, *125*(4), 1627–1682. doi: http://doi.org/10.1162/qjec.2010.125.4.1627

Rowe, D. C., Vesterdal, W. J., & Rodgers, J. L. (1998). Herrnstein's syllogism: Genetic and shared environmental influences on IQ, education, and income. *Intelligence*, *26*(4), 405–423. doi: http://doi.org/10.1016/S0160-2896(99)00008-2

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, public policy, and law*, *11*(2), 235. doi: http://doi.org/10.1037/1076-8971.11.2.235

Shriver, M. D., Parra, E. J., Dios, S., Bonilla, C., Norton, H., Jovel, C., ... others (2003). Skin pigmentation, biogeographical ancestry and admixture mapping. *Human genetics*, *112*(4), 387–399. doi: http://doi.org/10.1007/s00439-002-0896-y

Shuey, A. M. (1966). *The testing of negro intelligence* (2nd ed.). Social Science Press.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401–426. doi: http://doi.org/10.1016/j.intell.2006.09.004