Published in final edited form as:

Biodemography Soc Biol. 2014 ; 60(2): 212–228. doi:10.1080/19485565.2014.953029.

Data Quality Control in Social Surveys Using Genetic Information

Yi Li¹ and Guang Guo^{1,2,3}

¹Department of Sociology, University of North Carolina at Chapel Hill

²Carolina Population Center, University of North Carolina at Chapel Hill

³Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill

Abstract

This article introduces a novel way of taking advantage of genetic data in social surveys—for the purposes of data quality control. Genetic information could detect and repair data issues such as missing data, reporting errors, differences in measures of the same variable, and flawed data. Using data from two surveys, the College Roommate Study (ROOM) and the National Longitudinal Study of Adolescent Health (Add Health), we show that proportion identical by descent score (a measure of genetic relationships) can identify “misreported” and unreported sibling type, and detect misrepresented participants; bio-ancestry (a measure of ancestral population memberships) score can repair and recover missing race and discrepancies among different measures of self-reported race; and sex chromosomal information may help cross-check self-reported sex. This article represents an initial effort to utilize genetic data for the purposes of data quality control. As genetic data become increasingly available, researchers may explore more approaches to improve data quality.

INTRODUCTION

The past decade has seen a number of large-scale social surveys collect molecular genetic data, for example, the National Longitudinal Study of Adolescent Health (Harris, 2013), the Fragile Families Study (Reichman, Teitler, Garfinkel, & McLanahan, 2001), and the Health and Retirement Study (Crimmins et al., 2009). Recently, studies that incorporate genetic measurement have started to appear in leading journals (e.g., Caspi et al., 2002; Domingue, Fletcher, Conley, & Boardman, 2014; Mitchell et al., 2011; Shanahan, Vaisey, Erickson, & Smolen, 2008). Scholars in social and biological sciences are also calling for further integration of social, behavioral, and genetic research. In an effort toward this end, a national committee issued a report emphasizing the importance of research on interactions between social and genetic environmental factors and recommending development of efficient study designs and rigorous models to test gene and environment interactions over time (Hernandez & Blazer, 2006).

Genetic data not only enrich theoretical frameworks and empirical models in the social sciences, but also may offer promise with respect to data quality control in surveys. This article explores the role of genetic information in checking, repairing, and recovering self-reported variables. For example, when a self-reported variable has a missing value or is misreported, DNA data can provide additional information to detect and settle the issues in some circumstances. Proportion identical by descent (IBD) score (a measure of genetic relationships), bio-ancestry score (a measure of ancestral population memberships), and sex chromosomal information are all types of genetic information that may be used to improve self-reported variables such as sibling type, race, and sex. Improving data quality can be viewed a useful by-product of the availability of DNA data in social surveys.

In this article, we first describe the data sources and measures. Then we present the use of the proportion IBD score to check self-reported sibling type and to correct a more general issue—flawed data. Next, we show that missing race and discrepancies among different measures of self-reported race may in some cases be repaired by bio-ancestry score. Next, we demonstrate how to use sex chromosomal information to check self-reported sex. Lastly, we discuss implications of the findings and limitations.

DATA AND MEASURES

Data

Two data sources are employed in this article. The first is the College Roommate Study (ROOM) (Guo et al., 2009). ROOM was conducted at a public university in the spring semester of 2008 and sampled freshmen, sophomores, and juniors. Its main goal was to investigate peer influence on health-related behaviors such as drinking and smoking. Every year freshmen and returning students apply for housing. Students who do not request a specific roommate and do not request to participate in a themed housing program (e.g., substance free, foreign languages, health sciences, and global business) are eligible for randomly assigned roommates. These students were the targeted population. ROOM was a two-stage survey. The first stage was an online survey. Emails with the description of the survey and instructions on how to complete the survey were sent to students multiple times. Students were asked to answer questions about their health behaviors and socioeconomic backgrounds. The second stage involved collecting saliva from the students. Some 2,080 students (78.7 percent of those who completed the online survey) gave saliva to generate DNA data. Students who did not give saliva, lived off campus, or were studying abroad during the semester were eliminated for a final sample of 2,065 students.

The second data source is the National Longitudinal Study of Adolescent Health (Add Health). Add Health is a longitudinal study that surveyed a representative sample of U.S. adolescents in grades 7–12 during the 1994–95 school year (Harris, 2013). About 90,000 adolescents from 134 schools were surveyed. A subset of respondents, about 20,000, were interviewed at home and followed in subsequent waves. Interviewers asked for a wide range of information such as social backgrounds and physical health during interviews. Respondents were followed in 1996 (Wave II), 2001–02 (Wave III) and 2008–09 (Wave IV). The current article draws on data collected from a subsample of full siblings and twins at Wave III. At Wave I, a sibling subsample composed of full siblings, twins, half siblings, and

unrelated pairs, was drawn. At Wave III a total of 2,612 individuals in the sibling sample gave saliva. The final sample consists of 2,281 individuals whose DNA was successfully genotyped.

Measures of DNA

In ROOM participants provided saliva in an Oragene DNA collection kit (DNA Genotek; Ottawa, Ontario, Canada). DNA was extracted from 2 mls of saliva, with a median DNA yield of 27.3 µg. DNA was plated for Illumina genotyping at 30 µl at >50 ng/µl. In ROOM, we designed an Illumina GoldenGate assay for 384 candidate genetic variants—single nucleotide polymorphisms (SNPs)—in 28 genes (for technical details see Guo et al., 2009) and 350 SNPs in 28 genes were successfully genotyped. In Add Health, DNA gathered at Wave III in 2002 was isolated from buccal cells at the Institute of Behavior Genetics at the University of Colorado, Boulder. The average yield of DNA was 58 ± 1 µg. The genotype data used in this analysis were based on an Illumina GoldenGate assay. The GoldenGate array targeted 1,536 SNPs. A total of 1,140 SNPs in 130 genes were successfully genotyped. The 28 genes and 130 genes were selected mainly because of their implications in risky behaviors such as aggression, alcohol use, smoking, and illegal drug use.

Measures of Proportion Identical by Descent Score and Self-reported Sibling Type

Identical by descent (IBD) refers to alleles that are the same in pairs of persons, and are inherited from a common ancestor (Malécot, 1948; Wright, 1917). Proportion IBD score is a measure of how much a pair of persons is genetically linked to the same ancestor on a scale of 0 to 1. For example, the proportion IBD of identical or Monozygotic (MZ) twins is 1 because they share the same genome from the same parents. Fraternal or Dizygotic (DZ) twins and full siblings on average share 50 percent of the genome, for a proportion IBD score of 0.5 on average.

PLINK (Purcell et al., 2007), a whole genome association software program, estimates proportion IBD scores for all possible pairs in a dataset. PLINK uses the following formula to calculate proportion IBD scores:

$$P_1(IBM = 2) + 0.5 \times P_2(IBM = 1)$$

where P_1 is the proportion of loci possessing two identical alleles and P_2 is the proportion of the same loci possessing only one identical allele. The higher the score is, the more genome the pair shares. The value is 1 if two individuals have the same alleles, and 0 if they have no alleles in common.

Of the two surveys, only Add Health collected information on familial relationships. Respondents who had household members in the study were asked if they were twins or siblings. The types of pairs were further divided into MZ twins, DZ twins and full siblings. All mixed-sex twin pairs were classified as dizygotic. Same-sex twins were determined to be MZ or DZ based on self-reported confusability of appearance at Wave I (i.e., look like two peas in a pod as young children; and three other items, confused by strangers, teachers, or family members). When self-reported data on appearance was missing, the sibling type was

determined based on the mother's report of confusability of appearance. Some twin pairs were classified as uncertain status because the self-reported zygosity differs from the classification based on confusability of appearance. Add Health also collected survey information on parent-child pairs, but no genetic data on this type of pairs is available.

Of 2,281 individuals in the Add Health sample, there are 884 pairs, 1,768 individuals, in which both members have genotype data to obtain a proportion IBD score. Of the remaining 513 individuals, proportion IBD scores cannot be estimated for 501 because their siblings do not have genotype data. For the final 12, the scores cannot be estimated because information on whether the individuals are members of a sibling pair is not available.

Measures of Bio-ancestry Score and Self-reported Race

A panel of Ancestral Informative Markers (AIMs) is used to estimate bio-ancestry score (a measure of the genetic composition of geographic ancestral populations). AIMs are sets of genetic polymorphisms, whose allele frequencies differ significantly across populations (Frudakis et al., 2003; Parra et al., 1998; Shriver et al., 1997).

Both Add Health and ROOM targeted 186 SNPs as AIMs, which were developed at David Goldman's Laboratory of Neurogenetics at NIAAA for the purpose of detecting broad, continental African, East Asian, and European populations (Enoch, Shen, Xu, Hodgkinson, & Goldman, 2006). Of the 186 AIMs, 162 were successfully genotyped in ROOM, and 121 were successfully genotyped in Add Health. The AIMs are sufficient to distinguish African, East Asian and European populations, but are less effective in determining more specific sub-populations within the three populations (e.g., Chinese population within the East Asian population). Rosenberg and colleagues (2002) reported that 100 to 160 SNPs are sufficient when the sample size is about 1,000. In other studies similar conclusions were made (Halder, Shriver, Thomas, Fernandez, & Frudakis, 2008; Smith et al., 2001; Yang et al., 2005).

Bio-ancestry scores are estimated based on the AIM panel using STRUCTURE cluster analysis (Pritchard, Stephens, & Donnelly, 2000). STRUCTURE is a software package that investigates an individual's ancestral populations, independent of self-reported race. Assuming a mixture of multiple populations comprises the genome of each individual, the STRUCTURE analysis estimates the relative contributions of these populations to the individual. A K value needs to be set in the STRUCTURE analysis to indicate the hypothesized number of ancestral populations. The K contributions from K ancestry populations for each individual sum to 1.

Because our panel of AIMs was designed to differentiate continental populations of Africans, East Asians, and Europeans, we set $K=3$. In other words, individuals in the samples are assumed to have three bio-ancestry scores—African, East Asian, and European. The sum of the three scores is 1. To get more precise estimates, the STRUCTURE analysis is run 20 times and bio-ancestry scores are averaged on the 20 sets of results. All pair-wise symmetric similarity coefficients (SSC), which reflect the similarity of two sets of population structure estimates, are greater than 0.995. Our approach is similar to those in other studies of genetic population structure (Friedlaender et al., 2008; Wang et al., 2007).

ROOM data includes two indicators of self-reported race and ethnicity. The first is from a housing application that incoming freshmen submitted to university housing when requesting a dorm room. The university housing form only allowed students to self-identify as a member of one of six racial/ethnic groups: “White,” “Black,” “Hispanic,” “Asian and Pacific Islander,” “Native Indian” and “Other.” The second indicator comes from the online survey, which allowed respondents to mark one or more races. ROOM first asked respondents whether they were of Hispanic origin and then asked their race. Respondents can select all that apply from the five options: “White,” “Black,” “Asian and Pacific Islander,” “American Indian or Alaska Native” and “Other.”

Add Health collected multiple measures of race and ethnicity at Waves I and III, but not at Wave II. The Add Health race/ethnicity questionnaire first asked respondents to indicate whether they were of Hispanic origin. Next, respondents were allowed to identify with more than one racial group from “White,” “Black,” “Asian and Pacific Islander,” “American Indian or Alaska Native” and “Other.” In addition, a respondent was asked to indicate which single-race category best describes her if this respondent marked more than one race. During the interview at Wave III, interviewers were instructed to record what they thought to be the respondent’s single-best race (independent of what the respondent said). The categories available for interviewers were “White,” “Black,” “American Indian or Alaska Native,” “Asian or Pacific Islander” and “Other.” “Hispanic” was not an option.

PROPORTION IDENTICAL BY DESCENT SCORE, SIBLING TYPE, AND FLAWED DATA

Proportion Identical by Descent Score, Self-reported Sibling Type, and Identifying “Misreported” and Unreported Sibling Type

Table 1 presents proportion IBD score by sibling type in Add Health. Proportion IBD scores are estimated in PLINK using all of the available 1,140 SNPs. Results show that the mean proportion IBD scores for DZ twins and full siblings are both consistent with the expected value—0.5. However, the mean proportion IBD score for MZ twins is not 1.00, but 0.92. This is due to six pairs previously identified as MZ twins based on reported confusability. Proportion IBD scores of the six pairs range from 0.46 to 0.58. Thus, the six pairs of twins are DZ twins.

Proportion IBD score may also be used to help identify unreported or unknown sibling type. Five undermined twin pairs in Table 1 are classified according to the proportion IBD scores. One of five pairs is MZ twins with a proportion IBD score of 1.00. Four pairs are DZ twins with proportion IBD scores ranging from 0.04 to 0.59. Add Health classified some twin pairs of uncertain status on the basis of four variable number of tandem repeats (VNTR) loci and three tetra-nucleotide microsatellite loci. Our proportion IBD scores are consistent with results from the seven genetic markers.

Proportion Identical by Descent Score and Detecting Flawed Data

A more general usage of proportion IBD score is to detect flawed data such as misrepresented/falsified respondents. Sometimes respondents participate in a survey for a

second time under another person's name for incentives such as pay. It is also possible that due to errors in handling or data recording an individual's social and biological information may be recorded more than once. In these cases, proportion IBD score could help identify problematic pairs and correct them.

Case in point, in ROOM two participants were found to have a proportion IBD score of 1, meaning their DNA samples are identical. In the two identical DNA samples, 6 of 19 SNPs on X chromosomes are heterozygous. If SNPs on X chromosomes are heterozygous, it is very likely that the genetic sample belongs to a female. The genetic evidence, therefore, suggests that a female participant provided a DNA sample twice. Of the two participants, one reported as female in the online survey and the roommate was also female, whereas the other participant reported as male and the roommate was also male. Campus housing policies do not permit coed roommates, so the roommate sex confirms the reported sex of these respondents. Also, the female gave saliva first, and we would assume that if a person were providing a DNA sample in another person's name he or she would do it after providing his or her own DNA sample. Thus we concluded that the female may have given saliva twice for the monetary incentive. We decided not to link the male to this genetic sample. We also checked the female participant's online survey and found the answers to be within the reasonable range and no outliers that warranted exclusion. So we kept the survey responses.

BIO-ANCESTRY SCORE AND SELF-REPORTED RACE

Bio-ancestry Score and Self-reported Race

Tables 2 and 3 present average bio-ancestry scores for self-reported race in ROOM and Add Health respectively. For Add Health, results of Wave I and III are very close. We only present those based on Wave I self-reported race. Both Tables 2 and 3 show that bio-ancestry scores are closely associated with self-reported race. For example, in ROOM among self-reported "Black" in response to a multi-race question the average African ancestry is 0.89, and is 0.87 among blacks in response to a single-race question; the corresponding numbers in Add Health are 0.93 and 0.91 respectively.

When Self-reported Race Has Missing Values

Because bio-ancestry scores are closely associated with self-reported race, as shown in Tables 2 and 3, we propose that bio-ancestry scores can serve as a potential tool for data checking and repairing when reported race data are missing. However, missing or "misreported" race may carry sociological meanings. Depending on the goals of specific research, data users make the decision of whether to use bio-ancestry score to "impute" or replace reported race variables.

Table 4 presents bio-ancestry scores and housing race for all participants whose online survey race is missing in ROOM. The top panel of Table 4 shows that seven individuals' bio-ancestry scores are consistent with the race reported on their housing applications and the bottom panel shows that six individuals' bio-ancestry scores are not consistent with the housing application race. Note that in the bottom panel the six individuals' housing

application races are not “White,” “Black” or “East Asian.” If the race information in the housing applications was not available, race of these 13 respondents would be treated as missing. With the bio-ancestry scores, however, each of the 13 respondents could be assigned to a non-missing race variable.

Table 5 presents the bio-ancestry scores, self-reported race, and race observed by the interviewer for a representative subset of respondents whose self-reported race is missing in Add Health. Single-best race responses are not presented because they are consistent with multi-race responses for all these respondents. Table 5 is separated by the far right column into three panels. The top panel shows individuals whose bio-ancestry scores are consistent with the available race information, that is, self-reported race at another wave and interviewer-recorded race. In one of them (Pseudo ID is 1006), the interviewer-reported race is different from self-reported race. The respondent self-reported as “White” at Wave I but the interviewer marked the race to be “Asian” at Wave III. The European bio-ancestry score for this individual is 0.564, and Asian score is 0.312.

The middle panel shows individuals who are not classified as “White,” “Black” or “East Asian.” Bio-ancestry scores are still helpful in these cases. For example, the respondent with Pseudo ID 1012 reported as “Other” at Wave I, but the interviewer-recorded race at Wave III is “White.” The bio-ancestry scores indicate that this difference is not surprising, because this respondent has an African ancestry of 0.159, East Asian ancestry of 0.470, and European ancestry of 0.371. The bottom panel of Table 5 shows two individuals whose bio-ancestry is not consistent with the survey information. For both respondents, self-reported race and interviewer-recorded race are “White,” while the highest ancestry scores are East Asian.

Suppose that bio-ancestry were not available in Add Health. Although missing race could be saved by other race variables in certain circumstances, there are places in which reported race variables are not sufficient to resolve the problem. For respondents with Pseudo IDs 1006 and 1008 to 1014 in Table 5, bio-ancestry scores are needed for insight into why these discrepancies between the race variables exist. Take the case of Pseudo ID 1011 as an example, with “Other” race reported at Wave I and interviewer-reported race “White” at Wave III, it is difficult to decide which of the two variables should be put in the missing race at Wave III. In addition, for the case in which all of the self-reported races are missing (Pseudo ID 1005 in Table 5), the missing might be “imputed” as “White” by the interviewer-reported race. But with this person’s bio-ancestry scores, a clearer picture is seen as this person possesses about 49.9 percent European ancestry and 37.5 percent East Asian ancestry.

When Self-reported Race Differs across Different Measures or Waves

When self-reported race differs across various survey waves or measures, bio-ancestry score could offer additional information for reconciling such differences that might otherwise be considered reporting or recording errors.

Table 6 reports all the respondents who differ in bio-ancestry score, online race, or housing application race in ROOM. Yet, when comparing online survey race and housing application

race, the category “Hispanic” is excluded because of the different options participants were given. In the online survey participants were first asked whether they were “Hispanic” and then to choose from “White,” “Black,” “Asian and Pacific Islanders,” “Native Indian” or “Other.” On the housing application, however, participants were asked to choose only one race/ethnicity from “Hispanic,” “White,” “Black,” “Asian and Pacific Islanders,” “Native Indian” and “Other.”

Table 6 reveals that there are three types of differences in ROOM. “Misreporting” is the first, and is illustrated in the top panel. In this context, “misreporting” means a large gap is found between self-classified race and bio-ancestry. For example, the first responder (Pseudo ID 1001) has a very high African bio-ancestry score of 0.994, and reported as “Black” on the housing application, but indicated “White” in the online survey. The middle panel shows mixed-race respondents. Researchers have reported that mixed-race individuals are more likely to report different races (e.g., Hitlin, Brown, & Elder, 2006). Take the first respondent in the middle panel as an example (Pseudo ID 1003). This individual’s European ancestry score is 0.170 and African ancestry score is 0.817. With such a composition of bio-ancestry, this individual reported as “White” in online survey and “Black” on the housing application. The category for “White American Indian” is presented in the bottom panel of Table 6. All three respondents have very high European ancestry scores, ranging from 0.979 to 0.996, and they reported as “White” in the online survey. However, they all reported as “American Indian” on their housing applications. This finding may reflect “ethnic re-identification” among American Indians (e.g., Eschbach, 1993; Kelly & Nagel, 2002).

Table 7 presents respondents whose bio-ancestry scores, self-reported race at Waves I and III, and interviewer-identified race at Wave III do not match up in Add Health. For these respondents, single-best race is the same as multiracial-allowed race so only multiracial-allowed race is presented. In Add Health, there are two types of differences. “Misreporting” is the first. Two respondents in the top panel are in this category. They have very high Asian ancestry scores, 0.988 and 0.953 respectively, and both their Wave I race and interviewer-recorded race are “Asian.” But they reported as “White” at Wave III. In the bottom panel is the second type. Respondents in this category are mixed race according to their bio-ancestry scores and reported different races at different waves.

Without bio-ancestry scores it is almost impossible to distinguish “misreporting” and mixed race in Tables 6 and 7, and data users would have to treat the two groups the same. In Table 6, for example, two respondents (Pseudo IDs 1001 and 1003) have the same pattern—online race is “White” and housing race is “Black”. It turns out that the respondent with Pseudo ID 1001 has a high African ancestry of 0.994, but the respondent with Pseudo ID 1003 possesses an African ancestry of 0.817 and a European ancestry of 0.170.

THE SEX CHROMOSOMES AND SELF-REPORTED SEX

The Sex Chromosomes and Sex

Genetic information on sex is not as certain as that for genetic relationships and bio-ancestry due to the complex human sex-determining process and our limited understanding of it (Ellegren, 2011). XX and XY are the common combinations of sex chromosomes for

females and males respectively in humans. The gonads need to be formed and differentiated for humans to develop sexually. The Y chromosome induces testis formation and is the dominant determinant for male development. Without a Y chromosome, gonads differentiate into ovaries, resulting in female development. On the human Y chromosome, the Sex-determining Region of the Y chromosome (SRY) plays the crucial role. The SRY is a single exon gene. In this gene there are transcription initiation sites and these transcripts are identified in adult testis and other male tissues (Clepet et al., 1993). The SRY initiates testis development from early bipotential gonads (Koopman, Gubbay, Vivian, Goodfellow, & Lovell-Badge, 1991; Sinclair et al., 1990). The complex sex determination process also involves genes on autosomes. For example, *DMRT1*, which is located on chromosome 9, is found to be associated with haploinsufficiency in the form of XY sex reversal (Raymond et al., 1999).

Different combinations of the sex chromosomes other than XX and XY exist in the human population. XXY or XXYY males, known as Klinefelter syndrome, have two X chromosomes and one or two Y chromosomes. Also, persons with Turner syndrome have only one sex chromosome—that is, 45 chromosomes in total including only one X chromosome. The proportion of non-XX females and non-XY males is considerable. About 1 out of 500 to 1,000 boys are born with Klinefelter syndrome (Chen, 2005), and 1 of every 2,700 live births result in a child with Turner syndrome (Jaffe, 1999). Given the various forms of sex chromosome combinations, we maintain a cautious view on the application of genetic sex information in social surveys.

Due to the existence of different forms of sex chromosomes, it is impossible to be 100 percent certain of an individual's sex by only looking at the sex chromosomes. However, given the relatively low prevalence of unusual combination forms of sex chromosomes, the common forms of XX for females and XY for males may still serve as a reference.

Identifying “Misreported” Sex

In ROOM, 19 SNPs on X chromosomes and no SNPs on Y chromosomes were genotyped. So the genetic data is not entirely informative. X chromosomes are the only available information to suggest, rather than determine, biological sex.

One respondent who “misreported” sex is successfully identified using the sex chromosomal data in ROOM. The word “misreported” here refers to the situation in which self-reported sex differs from what biological sex and other information suggest. This respondent reported as male in the online survey but as female when applying for housing, and this person's roommate was female. It turns out 9 of 19 SNPs on the X chromosomes of this participant are heterozygous, which suggests a high probability of being female. Genetic sex, roommate sex, and housing application sex all indicate this is a female.

As mentioned previously, the sex chromosomal data also help identify misrepresented participant. In ROOM there is one pair of participants with identical genetic data, that is, with a proportion IBD score of 1. Information on the sex chromosomes suggests a high probability that the female in this pair gave saliva twice because 6 of 19 SNPs on the X chromosomes are heterozygous. The respondent's online survey sex and the roommate's sex

are both female (roommate sex should be concordant with ego sex because of the housing policies), confirming that this participant is female.

DISCUSSION AND CONCLUSION

Using data from ROOM and Add Health, we show that proportion IBD score can identify “misreported” and unreported sibling type and detect misrepresented participants; bio-ancestry score can in some circumstances help to save missing race, and reconcile discrepancies among different measures of self-reported race; and the sex chromosomes offer information that, coupled with other variables, is helpful in checking gender. In some cases, only with the genetic information can the issues be successfully detected and recovered. An example is that Table 6 shows that without the bio-ancestry scores it is almost impossible to distinguish between respondents who “misreported” race and those who are mixed-race.

The cost of using already-collected DNA data to implement quality control is minimal, but the benefits of doing so are considerable. A total number of 127 cases with various data issues can be repaired and recovered with the help of the genetic information: Six pairs of “misreported” MZ twins are found; five twin pairs of uncertain status are determined; one misrepresented respondent is corrected; 59 missing values in the race variables can be saved; four respondents who “misreported” their race are found; 40 discrepancies in the race variables are clarified; and one respondent who “misreported” sex is found. Particularly, for studies that analyze roommate pairs or sibling pairs, recovering one respondent means a pair can be saved. Moreover, tremendous amounts of financial support, time, personnel and other resources are invested in social surveys. If a large-scale survey spends \$100 million to collect data on 20,000 individuals, this represents an average of \$5,000 per participant. In this sense, it is worthwhile to repair every data issue possible.

We emphasize that genetic data are not a substitute for self-reported data. Sibling type, race and gender are all socially and biologically constructed in a complex way. The fact that respondents in ROOM and Add Health, most of whom are adolescents and young adults, change or do not report these variables at different waves should be understood from a sociological perspective in combination with other information. Social scientists have long been interested in racial and ethnic identities among adolescents and adults (e.g., Phinney, 1990). Identity development (Erikson, 1968), social identity theory (Tajfel & Turner, 1979) and acculturation theory (Park, 1928; Stonequist, 1937) provide theoretical frameworks for understanding why people may change racial and ethnic identities (for a review see Frable, 1997). From the social constructionist perspective gender identity is conceptualized as an ongoing performance. Individuals enact and reinforce their femininity/masculinity in social interactions (West & Zimmerman, 1987). Therefore, when a respondent “misreports” her race or gender on a questionnaire, it is possible that the respondent has changed her racial or gender identity. As shown in this article, with DNA data we can obtain more sociologically interesting information concerning race and gender.

The relatively small numbers of SNPs genotyped in ROOM (350) and Add Health (1,140) limit the findings of this article in several ways. First, IBD score is traditionally estimated

from pedigrees and the base population, which is the population that consists of the founders of the pedigrees. Ideally, with large SNP datasets IBD can be predicted without knowing pedigrees. However, because of the relatively small numbers of SNPs, the IBD score estimated here may not be as accurate as that obtained from dense SNP data. For the same reason, an alternative calculation of IBD score is used in this article. It should be noted that calculating IBD score can be more complicated than the calculation in this article (see Powell, Visscher, & Goddard, 2010 for a review on IBD). Second, the AIMs selected in this research detect three major continental populations; however we were not able to break down the African, East Asian and European classifications into more specific groups. Third, as for the sex chromosomes, only X chromosomes were genotyped and the numbers of genotyped SNPs on the X chromosomes are small. Given the complex nature of sex determination process, information about Y chromosomes and more SNPs on X chromosomes are desirable in the future.

This article provides an initial effort to improve survey data using genetic data. Rapid developments in genetics and the increasing collection of genetic data will introduce more ways to improve data quality. Recently, large-scale social surveys such as the Health and Retirement Study (Crimmins et al., 2009) started to release genotype data that contain millions of SNPs. The large datasets provide opportunities for further exploration of innovative methods for data quality control.

Acknowledgements

Two grants to Guang Guo supported the College Roommate Study (the William T. Grant Foundation) and the Illumina 1536 genotyping in Add Health (NSF's Human and Social Dynamics program BCS-0826913). Data from Add Health were funded by the National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies (www.cpc.unc.edu/addhealth/contract.html) to Kathleen Mullan Harris (P01-HD31921). Special acknowledgment is due Rick Bradley of the Housing Department, Kirk Wilhelmsen of the Genetics Department, Patricia Basta of the Bio-Specimen Process Center, Jason Luo of the Mammalian Genotyping Center, and the Odum Institute at the University of North Carolina, Chapel Hill. We received important assistance in SNP selection and the analysis of HGDP data from David Goldman and his Neurogenetics lab at NIAAA. Many hearty thanks go to Greg Duncan for his important role in the project. We are grateful to the Carolina Population Center (R24 HD050924) for general support. We thank Yilan Fu, Ted Mouw, Francois Nielsen, Michael Shanahan, Yang Yang and two anonymous reviewers for their helpful comments on earlier drafts of this article.

REFERENCES

- Caspi A, McClay J, Moffitt TE, Mill J, Martin J, Craig IW, . . . Poulton R. (2002). Role of Genotype in the Cycle of Violence in Maltreated Children. *Science*, 297(5582), 851–854. [PubMed: 12161658]
- Chen H (2005). Klinefelter Syndrome, 2014, from <http://emedicine.medscape.com/article/945649-overview#a0199>
- Clepet C, Schater AJ, Sinclair AH, Palmer MS, Lovell-Badge R, & Goodfellow PN (1993). The human SRY transcript. *Human Molecular Genetics*, 2(12), 2007–2012. [PubMed: 8111368]
- Crimmins EM, Guyer H, Langa KM, Ofstedal MB, Wallace RB, & Weir DR (2009). Documentation of Biomarkers in the Health and Retirement Study, 2014, from <http://hrsonline.isr.umich.edu/sitedocs/userg/HRS2006BiomarkerDescription.pdf>
- Domingue B, Fletcher J, Conley D, & Boardman J (2014). Genetic and educational assortative mating among US adults. *Proceedings of the National Academy of Sciences*, 111(22), 7996–8000.
- Ellegren H (2011). Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. *Nature Reviews Genetics*, 12, 157–166.
- Enoch M, Shen P, Xu K, Hodgkinson C, & Goldman D (2006). Using ancestry-informative markers to define populations and detect population stratification. *Journal of Psychopharmacology*, 20, 19–26.

- Erikson EH (1968). *Identity: Youth and Crisis*. New York: Norton.
- Eschbach K (1993). Changing Identification among American Indians and Alaska Natives. *Demography*, 30(4), 635–652. [PubMed: 8262285]
- Frale DES (1997). Gender, Racial, Ethnic, Sexual, and Class Identities. *Annual Review of Psychology*, 48, 139–162.
- Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, ... Weber JL (2008). The genetic structure of Pacific islanders. *Plos Genetics*, 4, 173–190.
- Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Gijupalli S, Gunturi S, ... Nachimuthu PK (2003). A classifier for the SNP-Based inference of ancestry. *Journal of Forensic Sciences*, 48, 771–782. [PubMed: 12877291]
- Guo G, Hardie JH, Owen C, Daw JK, Fu Y, Lee H, ... Duncan G (2009). DNA Collection in a Social Science Study: A Pilot Study of Peer Impacts on Attitudes and Drinking Behavior. *Sociological Methodology*, 39, 1–29.
- Halder I, Shriver M, Thomas M, Fernandez JR, & Frudakis T (2008). A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: Utility and applications. *Human Mutation*, 29, 648–658. [PubMed: 18286470]
- Harris KM (2013). Design Features of Add Health, 2014, from <http://www.cpc.unc.edu/projects/addhealth/data/guides/DesignPaperWIIV.pdf>
- Hernandez LM, & Blazer DG (2006). *Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture*. Washington, DC: National Academies Press.
- Hitlin S, Brown JS, & Elder GH (2006). Racial self-categorization in adolescence: Multiracial development and social pathways. *Child Development*, 77(5), 1298–1308. [PubMed: 16999799]
- Jaffe RB (1999). Disorders of Sexual Development. In Yen SSC, Jaffe RB & Barbieri RL (Eds.), *Reproductive endocrinology: physiology, pathophysiology, and clinical management*: Philadelphia: W.B. Saunders.
- Kelly ME, & Nagel J (2002). Ethnic re-identification: Lithuanian Americans and Native Americans. *Journal of Ethnic and Migration Studies*, 28(2), 275–289.
- Koopman P, Gubbay J, Vivian N, Goodfellow P, & Lovell-Badge R (1991). Male Development of Chromosomally Female Mice Transgenic for *Sry*. *Nature*, 351, 117–121. [PubMed: 2030730]
- Malécot G (1948). *Les Mathématiques de L'Hérédité*. Paris: Masson & Cie.
- Mitchell C, Notterman D, Brooks-Gunn J, Hobcraft J, Garfinkel I, Jaeger K, ... McLanahan S (2011). Role of mother's genes and environment in postpartum depression. *Proceedings of the National Academy of Sciences*, 108(20), 8189–8193.
- Park R (1928). Human Migration and the Marginal Man. *American Journal of Sociology*, 33(6), 881–893.
- Parra EJ, Marcini A, Akey L, Martinson J, Batzer MA, Cooper R, ... Shriver MD (1998). Estimating African American admixture proportions by use of population-specific alleles. *American Journal of Human Genetics*, 63, 1839–1851. [PubMed: 9837836]
- Phinney JS (1990). Ethnic Identity in Adolescents and Adults: Review of Research. *Psychological Bulletin*, 108(3), 499–514. [PubMed: 2270238]
- Powell JE, Visscher PM, & Goddard ME (2010). Reconciling the analysis of IBD and IBS in complex traits studies. *Nature Reviews Genetics*, 11, 800–805.
- Pritchard JK, Stephens M, & Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959. [PubMed: 10835412]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, David Benderb a., Julian Maller, ... Sham PC (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, 81(3), 559–575. [PubMed: 17701901]
- Raymond CS, Parker ED, Kettlewell JR, Brown LG, Page DC, Kusz K, ... Zarkower D (1999). A region of human chromosome 9p required for testis development contains two genes related to known sexual regulators. *Human Molecular Genetics*, 8(6), 989–996. [PubMed: 10332030]
- Reichman NE, Teitler JO, Garfinkel I, & McLanahan SS (2001). Fragile Families: sample and design. *Children and Youth Services Review*, 23, 303–326.

- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, & Feldman MW (2002). Genetic structure of human populations. *Science*, 298, 2381–2385. [PubMed: 12493913]
- Shanahan M, Vaisey S, Erickson LD, & Smolen A (2008). Environmental Contingencies and Genetic Propensities: Social Capital, Educational Continuation, and Dopamine Receptor Gene DRD2. *American Journal of Sociology*, 114(s1), s260–s286.
- Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, & Ferrell RE (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*, 60, 957–964. [PubMed: 9106543]
- Sinclair AH, Berta P, Palmer MS, Hawkins JR, Griffiths BL, Smith MJ, ... Goodfellow PN (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature*, 346(6281), 240–244. [PubMed: 1695712]
- Smith MW, Lautenberger JA, Shin HD, Chretien JP, Shrestha S, Gilbert DA, & O'Brien SJ (2001). Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. *American Journal of Human Genetics*, 69, 1080–1094. [PubMed: 11590548]
- Stonequist E (1937). *The Marginal Man: A Study in Personality and Culture Conflict*. New York: Scribner's Sons.
- Tajfel H, & Turner J (1979). An integrative theory of intergroup conflict In Austin W & Worchel S (Eds.), *The Social Psychology of Intergroup Relations*. Monterey, CA: Brooks/Cole.
- Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, ... Ruiz-Linare A (2007). Genetic variation and population structure in Native Americans. *Plos Genetics*, 3, 2049–2067.
- West C, & Zimmerman D (1987). *Doing Gender*. *Gender and Society*, 1, 125–151.
- Wright S (1917). Coefficients of inbreeding and relationship. *American Naturalist*, 56, 330–338.
- Yang N, Li HZ, Criswell LA, Gregersen PK, Alarcon-Riquelme ME, Kittles R, ... Seldin MF (2005). Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Human Genetics*, 118, 382–392. [PubMed: 16193326]

TABLE 1

Proportion IBD Score by Reported Sibling Pair Type, Add Health

Reported Sibling Pair Type	Proportion IBD Score		N of Pairs
	Mean	Std	
Full Sibling	0.52	0.11	633
DZ Twins	0.51	0.13	212
MZ Twins	0.92 ^a	0.18	34
Undetermined Twins	0.48	0.35	5

^aProportion IBD scores of six pairs are not 1.00. The six pairs were determined to be MZ twins by self-reported confusability. Proportion IBD scores suggest the six pairs are DZ twins.

TABLE 2

Distribution of Average Bio-ancestry Score for Self-reported Race, ROOM

	Average Bio-ancestry Score			
Self-reported Race	European	African	East Asian	Sample Size
Multiracial-allowed Race (Online Race)				
White	0.98	0.01	0.01	1,406
Black	0.09	0.89	0.02	281
East Asian	0.04	0.00	0.96	86
South Asian	0.68	0.05	0.26	41
American Indian	0.64	0.16	0.20	4
Multiracial	0.61	0.26	0.13	172
Other	0.69	0.15	0.15	62
Missing	0.81	0.17	0.01	13
Total				2,065
Single Race (Housing Application Race)				
White	0.98	0.01	0.01	1,331
Black	0.11	0.87	0.02	338
Asian	0.29	0.02	0.69	151
American Indian	0.80	0.09	0.11	18
Hispanic	0.82	0.07	0.11	132
Other	0.84	0.09	0.07	94
N/A	0.98	0.02	0.01	1
Total	0.77	0.16	0.07	2,065

TABLE 3

Distribution of Average Bio-ancestry Score for Self-reported Race, Add Health

	Average Bio-ancestry Score			
Self-reported Race	European	African	East Asian	Sample Size
Multiracial-allowed Race, Wave I				
White	0.95	0.02	0.03	1,437
Black	0.06	0.93	0.01	381
Asian	0.07	0.01	0.92	160
American Indian	0.63	0.05	0.32	19
Multiracial	0.67	0.21	0.12	92
Other	0.61	0.10	0.29	179
Missing	0.53	0.39	0.08	13
Total				2,281
Single Best Race, Wave I				
White	0.95	0.02	0.03	1,486
Black	0.07	0.91	0.02	406
Asian	0.10	0.01	0.89	171
American Indian	0.66	0.06	0.28	24
Other	0.61	0.10	0.28	180
Missing	0.53	0.39	0.08	14
Total				2,281

TABLE 4

Bio-ancestry Score and Housing Application Race for All Respondents Whose Online Survey Race is Missing, ROOM

Pseudo ID	Bio-ancestry Score			Housing Application Race	Bio-ancestry Consistent with Housing Application Race?
	European	African	East Asian		
1001	0.992	0.003	0.005	White	Yes
1002	0.996	0.001	0.002	White	Yes
1004	0.996	0.003	0.002	White	Yes
1005	0.976	0.002	0.022	White	Yes
1006	0.992	0.002	0.005	White	Yes
1007	0.995	0.002	0.003	White	Yes
1003	0.178	0.820	0.002	Black	Yes
1008	0.845	0.136	0.019	Hispanic	No
1009	0.672	0.270	0.059	Hispanic	No
1010	0.985	0.012	0.003	Hispanic	No
1011	0.993	0.002	0.005	Other	No
1012	0.963	0.028	0.010	Other	No
1013	0.009	0.976	0.014	Other	No

Bio-ancestry Score, Self-reported Race, and Interviewer-reported Race for a Representative Subset of Respondents whose Self-reported Race is Missing, Add Health

TABLE 5

Pseudo ID	Bio-ancestry Score			Self-reported Race		Interviewer-reported Race		Bio-ancestry Consistent with Available Race?
	European	African	East Asian	Wave I	Wave III	Wave III	Wave III	
1001	0.978	0.008	0.014	Missing	White	White	White	Yes
1002	0.616	0.071	0.313	Missing	White	White	White	Yes
1003	0.050	0.902	0.048	Missing	Black	Black	Black	Yes
1004	0.002	0.997	0.001	Missing	Black	Black	Black	Yes
1005	0.499	0.126	0.375	Missing	Missing	Missing	White	Yes
1006	0.564	0.124	0.312	White	Missing	Missing	Asian	Yes
1007	0.995	0.003	0.002	White	Missing	Missing	White	Yes
1008	0.642	0.110	0.248	American Indian	Missing	Missing	Asian	No
1009	0.279	0.006	0.715	American Indian	Missing	Missing	White	No
1010	0.590	0.048	0.362	American Indian	Missing	Missing	White	No
1011	0.980	0.016	0.004	Other	Missing	Missing	White	No
1012	0.371	0.159	0.470	Other	Missing	Missing	White	No
1013	0.647	0.023	0.331	Other	Missing	Missing	American Indian	No
1014	0.554	0.444	0.003	Multiracial	Multiracial	Multiracial	Black	No
1015	0.386	0.172	0.442	White	Missing	Missing	White	No
1016	0.446	0.005	0.550	White	Missing	Missing	White	No

TABLE 6

All Respondents Who Differ among Bio-ancestry Score, Online Survey Race and Housing Application Race, ROOM

Bio-ancestry Score					
Pseudo ID	European	African	East Asian	Online Race	Housing Application Race
“Misreporting” ^a					
1001	0.004	0.994	0.002	White	Black
1002	0.992	0.002	0.006	Black	White
Mixed Race					
1003	0.170	0.817	0.013	White	Black
1004	0.865	0.006	0.129	White	Asian
1005	0.363	0.003	0.635	White	Other
1006	0.520	0.476	0.004	Black	Black
1007	0.824	0.165	0.011	Black	Black
1008	0.608	0.386	0.006	Multiracial	Black
1009	0.539	0.458	0.002	Multiracial	Black
1010	0.612	0.381	0.007	Multiracial	Black
1011	0.590	0.406	0.004	Multiracial	Black
1012	0.448	0.003	0.549	Multiracial	White
1013	0.851	0.005	0.144	Multiracial	Asian
1014	0.516	0.004	0.48	Multiracial	Asian
1015	0.507	0.009	0.484	Multiracial	Asian
1016	0.533	0.003	0.464	Multiracial	Asian
1017	0.783	0.011	0.207	Multiracial	Asian
1018	0.585	0.008	0.407	Multiracial	Asian
1019	0.508	0.003	0.489	Multiracial	Asian
1020	0.573	0.005	0.421	Multiracial	Asian
1021	0.762	0.023	0.214	Multiracial	Asian
1022	0.503	0.002	0.496	Multiracial	Asian
1023	0.493	0.015	0.492	Multiracial	Asian
1024	0.501	0.494	0.005	Other	Black
1025	0.542	0.064	0.394	Other	Black
“White American Indian”					
1026	0.979	0.003	0.018	White	American Indian
1027	0.990	0.002	0.008	White	American Indian
1028	0.996	0.002	0.002	White	American Indian

^a“Misreporting” in this context means a large gap exists between the respondent’s self-reported race and bio-ancestry score.

TABLE 7
All Respondents Who Differ among Bio-ancestry Score, Self-reported Race and Interviewer-reported Race, Add Health

Pseudo ID	Bio-ancestry Score			Self-reported Race		Interviewer-reported Race	
	European	African	East Asian	Wave I	Wave III	Wave III	Wave III
“Misreporting” ^a							
1001	0.008	0.004	0.988	Asian	White		Asian
1002	0.035	0.012	0.953	Asian	White		Asian
Mixed Race							
1003	0.863	0.119	0.018	White	Black		White
1004	0.843	0.030	0.127	White	Asian		American Indian
1005	0.657	0.325	0.018	Black	Black		Black
1006	0.542	0.437	0.021	Black	Black		Black
1007	0.400	0.585	0.015	Black	Asian		American Indian
1008	0.420	0.053	0.527	Asian	White		Asian
1009	0.270	0.117	0.614	American Indian	White		White
1010	0.315	0.061	0.624	American Indian	White		American Indian
1011	0.341	0.004	0.654	Other	White		White
1012	0.475	0.010	0.515	Other	White		White
1013	0.386	0.070	0.544	Other	White		White
1014	0.384	0.120	0.496	Other	White		White
1015	0.358	0.202	0.440	Other	White		White
1016	0.347	0.173	0.481	Other	White		White
1017	0.249	0.389	0.361	Other	White		White
1018	0.023	0.341	0.636	Other	Black		Black
1019	0.828	0.153	0.020	Other	Black		Black

^a“Misreporting” in this context means a large gap exists between the respondent’s self-reported race and bio-ancestry score.