Published in final edited form as:
Genet Med. 2010 October ; 12(10): 648–650. doi:10.1097/GIM.0b013e3181efe2df.

Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records

Logan Dumitrescu, BS[1,2], **Marylyn D. Ritchie, PhD, MS**[1,2], **Kristin Brown-Gentry, MS**[2], **Jill M. Pulley, MBA**[3], **Melissa Basford, MBA**[3], **Joshua C. Denny, MD, MS**[4,5], **Jorge R. Oksenberg, PhD**[6], **Dan M. Roden, MD**[3,5,7], **Jonathan L. Haines, PhD**[1,2], and **Dana C. Crawford, PhD**[1,2,*]

[1] Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN

[2] Center for Human Genetics Research, Vanderbilt University, Nashville, TN

[3] Office of Personalized Medicine, Vanderbilt University, Nashville, TN

[4] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN

[5] Department of Medicine, Vanderbilt University, Nashville, TN

[6] Department of Neurology, University of California at San Francisco, San Francisco, CA

[7] Department of Pharmacology, Vanderbilt University, Nashville, TN

Abstract

Purpose—The Vanderbilt DNA Databank (BioVU) is a biorepository that currently contains >80,000 DNA samples linked to electronic medical records. While BioVU is a valuable source of samples and phenotypes for genetic association studies, it is unclear whether the administratively assigned race/ethnicity in BioVU can accurately describe and be used as a proxy for genetic ancestry.

Methods—We genotyped 360 SNPs on the Illumina DNA Test Panel containing ancestry informative markers (AIMs) in 1910 BioVU samples with observer-reported ancestry and 384 samples from the Multiple Sclerosis Genetics Group with self-reported ancestry. Genetic ancestry was inferred for all individuals using STRUCTURE 2.2.

Results—More than 98% of observer-reported European Americans (EA) was genetically inferred to have at least 60% European ancestry. Ninety-three of observer-reported African Americans (AA) was genetically inferred to be predominantly of African ancestry. We determined that the concordance of observer-reported race/ethnicity and inferred genetic ancestry was not significantly different from that of self-reported race/ethnicity in either population (p=0.09 and 0.94 in European Americans and African Americans, respectively).

Conclusions—Observer-reported race/ethnicity for European Americans and African Americans approximates genetic ancestry as well as self-reported race/ethnicity, making biorepositories linked to EMRs such as BioVU a viable source of DNA samples for future large-scale genetic association studies.

Keywords

biorepositories; admixture; ancestry; electronic medical record; population stratification

[*]Corresponding Author: Dana C. Crawford, Center for Human Genetics Research, Vanderbilt University 2215 Garland Avenue, 515B Light Hall Nashville, TN 37232, Phone: 615-343-7852, Fax: 615-322-6974, crawford@chgr.mc.vanderbilt.edu.

INTRODUCTION

Ancestry (or genetic background) information is imperative for proper genetic association study design and for control of population stratification[1–3]. If case and control samples are drawn from dissimilar ancestral populations, significant associations may actually represent underlying genetic differentiation among samples and not associations with the phenotype under study[4–7]. Hundreds to thousands of markers across the genome can be used to estimate genetic ancestry and to adjust for population stratification in downstream tests of association. Common sources of these data in today's climate of high-throughput, cost effective genotyping are data from genome-wide association studies (GWAS) or standard panels of ancestry informative markers (AIMs)[8,9]. Despite the wide availability of these data and genotyping assays, there remain many large genetic association studies that do not yet have either GWAS or AIMs data available for all or most DNA samples in the dataset[10–12]. As such, these genetic association studies still rely on self-reported race/ethnicity as a proxy for genetic ancestry as most data support the assumption that self-reported race/ethnicity approximates genetic ancestry, particularly for samples with substantial genetic differentiation[13–15].

While self-reported race/ethnicity is common in genetic association studies, many clinic and hospital-based studies use observer or interviewer reported ancestry rather than self-report. The concordance between race/ethnicity recorded in the medical charts and clinical databases and self-reported race/ethnicity has been examined in recent years[16–18]. In one small exploratory study, 22–33% of respondents with diverse racial/ethnic backgrounds viewed themselves differently than how they were categorized in a community health center database[17]. In contrast, another report of Veterans Affairs healthcare users found that observer-reported race/ethnicity agreed with self-reported race/ethnicity for most users (95%)[18]. Thus, the concordance between observer-reported and self-described race/ethnicity varies, and this variation is most likely dependent on the methods of interview, site of study, and the racial/ethnic composition of the population under study.

The Vanderbilt DNA Databank (BioVU) is a biorepository of >80,000 DNA samples linked to electronic medical records (EMR) in Nashville, TN. BioVU uses discarded blood samples collected during routine patient care[19]. DNA is extracted and linked to de-identified data obtained and routinely updated from the EMR and other administrative databases. This approach has the advantage of scale, enabling genotype-phenotype associations across a variety of clinical outcomes represented in the patient population[10,19]. While BioVU is a valuable source of DNA samples and phenotypes for genetic association studies, it is unclear whether race/ethnicity, which is administratively assigned in BioVU, can be used as a proxy for ancestry in future genetic association studies in the absence of high density genotype data across the genome (such as GWAS data) or incurring the cost of additional genotyping of AIMs.

BioVU does provide observer-reported race/ethnicity, but a report of self-identified race/ethnicity is not available for direct comparison; thus, an alternative approach was necessary to explore the possible differences between the two types of reporting methods. To assess the use of observer-reported race/ancestry as a proxy for ancestry in BioVU-based genetic association studies, we genotyped 360 markers on the Illumina DNA Test Panel which includes ancestry informative markers (AIMs) in a subset of BioVU samples (observer-reported race/ethnicity data) and in a sample ascertained by the Multiple Sclerosis Genetics Group (self-reported race/ethnicity data) to infer genetic ancestry. The percent concordance of reported and inferred genetic ancestry was calculated in each group separately. We then tested for differences between the concordance of observer-reported race/ethnicity with inferred genetic ancestry and the concordance of self-reported ancestry with inferred genetic ancestry. Results of these comparisons demonstrate that observer-reported race/ethnicity in BioVU approximates inferred genetic ancestry as well as self-reported race/ethnicity, suggesting that observer-

reported race/ethnicity recorded in BioVU is an acceptable proxy for genetic ancestry for most DNA samples studied.

METHODS AND MATERIALS

Study Populations

A full description of the BioVU resource and its ethical protections has been described elsewhere19. A subset of BioVU samples was used in these analyses (n=1,910). The Multiple Sclerosis Genetics Group (MSGG) was founded in 1989 to study the role heredity plays in multiple sclerosis; consent and ascertainment are detailed elsewhere[20]. A random subset of unrelated controls from the MSGG was used for these analyses (n=384).

Genotyping

Both BioVU and MSGG samples were genotyped using the Illumina DNA Test Panel, which contains 360 validated single nucleotide polymorphisms (SNPs) distributed across the genome (Supplementary Table 1). All genotyping was performed on the Illumina BeadXpress[21]. For the majority of these SNPs, allele frequencies differ greatly between the major HapMap populations and thus can be used as ancestry informative markers (AIMs).

Prior to analysis, SNPs were filtered to exclude those with low minor allele frequency (MAF<1%), deviations from Hardy-Weinberg expectations ($p<10^{-4}$), and low genotyping efficiency (<95%). A total of 294 and 341 SNPs in the samples with self-reported and observer-reported ancestry, respectively, were analyzed.

Statistical Methods

For this analysis we focused on two racial/ethnic groups, European Americans (EA) and African Americans (AA), as these two groups represent the majority of BioVU samples (78.7% and 10.5%, respectively). For each study population (BioVU and MSGG) and each racial/ethnic group (EA and AA), the proportion of samples whose genetic ancestry (inferred by STRUCTURE 2.222) matched their reported ancestry was calculated. A two-sample test of proportion was used to test for differences between the concordance of observer-reported race/ethnicity with inferred genetic ancestry and the concordance of self-reported race/ethnicity with inferred genetic ancestry. Statistical significance was defined as $p<0.05$.

RESULTS

Observer-reported European Americans represent the majority of BioVU DNA samples in this study (78.7%). Of these 1,503 samples, 1,481 (98.5%) were inferred as having predominantly (>60%) European ancestry, including 1,439 (95.7%) participants with greater than 90% European ancestry (Table 1). Of the self-reported European Americans, all samples were inferred to have at least 75% European ancestry. This concordance rate was not significantly different than that calculated from observer-reported European Americans when lower ancestry proportion thresholds were used (p=0.10, 50% threshold; p=0.09, 60% threshold). When the threshold for classification was increased to greater than 75%, the difference in concordance with inferred genetic ancestry between self- and observer-report became statistically significant (p=0.04). At the strictest threshold of 90%, the percent concordance of observer-reported race was significantly higher than that of self-report (95.7% versus 88.0%, respectively; $p<0.001$).

The second most prominent ethnic/racial group in BioVU is African Americans (10.5%). Observer-reported race/ethnicity was able to distinguish samples with inferred African genetic ancestry from those of non-African genetic ancestry; however, the resulting inferred African

ancestry samples were not as homogenous as the inferred European ancestry samples. That is, of the 201 observer-reported African Americans, 187 (93.0%) were of predominantly African genetic ancestry, but only 44 (21.9%) were inferred to have greater than 90% African ancestry (Table 1). This distribution is not unexpected given the amount of admixture inherent in African American populations. Moreover, the concordance of observer-reported race/ethnicity with inferred genetic ancestry was not significantly different from the concordance of self-reported race/ethnicity in African Americans, regardless of the stringency of the threshold ($p>0.34$ at all four thresholds).

DISCUSSION

Our data indicate that observer-reported race/ethnicity in BioVU can be used as a proxy for genetic ancestry. We found a high concordance between observer-reported race/ethnicity and genetic ancestry, especially in European Americans. Furthermore, we determined that observer-reported race/ethnicity has a similar percent concordance with genetically defined ancestry as that of the self-reported race/ethnicity, which is widely used and accepted as a proxy for ancestry in genetic epidemiology.

We acknowledge, however, that this proxy is imperfect and that this imperfection may, in part, reflect variability in observer reports, a variable we are not able to quantify easily in BioVU. Also, we must remain cautious in our interpretation when considering observer-reported African Americans, a historically admixed population. While the majority of observer-reported European Americans fell into one genetic cluster, observer-reported African Americans had a broader distribution of percent genetically inferred ancestry. This distribution has been observed in many studies of admixture in African Americans[14,23–26]. Therefore, while reported race/ethnicity is able to categorize individuals as having a majority of European or African genetic ancestry, it cannot estimate or account for the admixture inherent in populations of African descent in the United States.

Our study focused on European and African Americans, since they are the predominant racial/ethnic groups in BioVU. However, it would be beneficial to expand our analysis to a larger representation of individuals with other observer-reported race/ethnicities (e.g. Hispanics, Asians, Native Americans, and Other). To date, greater than 1,300 (~2%) of BioVU samples fall into this category. However, use of HapMap samples as pseudo-ancestrals to infer genetic ancestry may prove problematic for these groups given that none of the major HapMap populations are perfect proxies for Native Americans, Hispanics, or Asians (not of eastern Asian descent). The recent expansion of populations available in HapMap may alleviate this problem, but further studies are needed to assess the utility of these additional HapMap populations compared with an outbred, diverse population that is characteristic of the United States. Also, this study did not address the small percentage of records for which observer race/ethnicity is absent. Our prior studies suggest this set of missing race/ethnicity is small, varying between 3–9%[10,27]. Use of automated methods to extract references to race/ethnicities from clinical notes may prove beneficial to fill these gaps in data.

Determining the feasibility of using observer-reported race/ethnicity as a proxy for genetic ancestry is crucial for all future genetic association studies using biorepositories linked to EMRs such as BioVU. In support of our conclusions, recent tests of association using observer-reported European-descent cases and controls in BioVU suggest that this proxy for genetic ancestry is sufficient to replicate well-known GWAS and candidate gene associations[10]. Further studies, however, are needed to determine if the observations reported here are true for other biorepositories linked to EMRs given differences among administrative and demographic data collections in clinical settings across the United States. Nevertheless, we demonstrate that observer-reported race/ethnicity for European Americans and African Americans

approximates genetic ancestry as well as self-reported race/ethnicity, suggesting biorepositories based on EMRs such as BioVU may be a viable source of DNA samples for future large-scale genetic association studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported, in part, by the National MS Society grants RG3060 and RG2899 (J.R. Oksenberg), VICTR (the Vanderbilt Institute for Clinical and Translational Research; CTSA grant 1UL1 RR024975-01 from NCRR/NIH), which also provides partial support for BioVU, and NIH grant HG004603. The Vanderbilt DNA Resources Core housed the DNA samples and also performed the genotyping for this work. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work. All genotyping was performed by the Vanderbilt DNA Resources Core.

Reference List

1. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet 2003;361:598–604. [PubMed: 12598158]
2. Koller DL, Peacock M, Lai D, Foroud T, Econs MJ. False positive rates in association studies as a function of degree of stratification. J Bone Miner Res 2004;19:1291–1295. [PubMed: 15231016]
3. Lander ES, Botstein D. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 1986;51(Pt 1):49–62. [PubMed: 2884068]
4. Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55:997–1004. [PubMed: 11315092]
5. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220–8. [PubMed: 10364535]
6. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet 2000;67:170–81. [PubMed: 10827107]
7. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959. [PubMed: 10835412]
8. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet 2008;17:R143–R150. [PubMed: 18852203]
9. Kosoy R, Nassir R, Tian C, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 2009;30:69–78. [PubMed: 18683858]
10. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010;86:560–572. [PubMed: 20362271]
11. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): Design, methods and recruitment for a large population-based biobank. Personalized Med 2005;2:49–79.
12. Design and estimation for the National Health Interview Survey, 1995-2004. Vital Health Stat 2000;2:1–31.
13. Burchard EG, Ziv E, Coyle N, et al. The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 2003;348:1170–1175. [PubMed: 12646676]
14. Sinha M, Larkin EK, Elston RC, Redline S. Self-reported race and genetic admixture. N Engl J Med 2006;354:421–422. [PubMed: 16436780]
15. Tang H, Quertermous T, Rodriguez B, et al. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 2005;76:268–275. [PubMed: 15625622]
16. Cho MK. Racial and ethnic categories in biomedical research: there is no baby in the bathwater. J Law Med Ethics 2006;34:497–9. 479. [PubMed: 17144171]

17. Moscou S, Anderson MR, Kaplan JB, Valencia L. Validity of racial/ethnic classifications in medical records data: an exploratory study. Am J Public Health 2003;93:1084–1086. [PubMed: 12835189]

18. Sohn MW, Zhang H, Arnold N, et al. Transition to the new race/ethnicity data collection standards in the Department of Veterans Affairs. Popul Health Metr 2006;4:7. [PubMed: 16824220]

19. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008;84:362–369. [PubMed: 18500243]

20. Cree BA, Khan O, Bourdette D, et al. Clinical characteristics of African Americans vs Caucasian Americans with multiple sclerosis. Neurology 2004;63:2039–2045. [PubMed: 15596747]

21. Lin CH, Yeakley JM, McDaniel TK, Shen R. Medium- to high-throughput SNP genotyping using VeraCode microbeads. Methods Mol Biol 2009;496:129–142. [PubMed: 18839109]

22. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959. [PubMed: 10835412]

23. Bryc K, Auton A, Nelson MR, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A 2010;107:786–791. [PubMed: 20080753]

24. Tishkoff SA, Reed FA, Friedlaender FR, et al. The genetic structure and history of Africans and African Americans. Science 2009;324:1035–1044. [PubMed: 19407144]

25. Lind JM, Hutcheson-Dilks HB, Williams SM, et al. Elevated male European and female African contributions to the genomes of African American individuals. Hum Genet 2007;120:713–722. [PubMed: 17006671]

26. Smith MW, Patterson N, Lautenberger JA, et al. A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet 2004;74:1001–1013. [PubMed: 15088270]

27. Bastarache L, Ritchie Marylyn D, Crawford Dana C, Basford MA, Denny Joshua C. Detecting race and ethnicity using natural language processing. Proc AMIA Symp. 2009

Table 1

Comparison of the concordance rate of self-reported race/ethnicity and observer-reported race/ethnicity in European Americans and African Americans

Genetic ancestry was inferred using STRUCTURE 2.2. Assuming K=3 and allowing 50,000 iterations and 10,000 burn-in cycles, ancestry proportions were determined for all samples using unrelated individuals from HapMap Phase 3 (109 CEU, 108 YRI, and 164 JPT/CHB) as learning populations for STRUCTURE. Four different thresholds (50%, 60%, 75%, and a most stringent 90%) were used to classify the inferred genetic ancestry of each sample.

	Percent concordance with inferred genetic ancestry		p-value
	Self-reported race/ethnicity	Observer-reported race/ethnicity	
European Americans			
50% Threshold	100% (192/192)	98.6% (1,482/1,503)	0.10
60% Threshold	100% (192/192)	98.5% (1,481/1,503)	0.09
75% Threshold	100% (192/192)	97.9% (1,471/1,503)	0.04
90% Threshold	88.0% (169/192)	95.7% (1,439/1,503)	<0.001
African Americans			
50% Threshold	97.9% (188/192)	96.5% (194/201)	0.40
60% Threshold	93.2% (179/192)	93.0% (187/201)	0.94
75% Threshold	74.0% (142/192)	76.1% (153/201)	0.63
90% Threshold	26.0% (50/192)	21.9% (44/201)	0.34