HUGO J
DOI 10.1007/s11568-011-9154-5

RESEARCH ARTICLE

The use of race, ethnicity and ancestry in human genetic research

Sarah E. Ali-Khan · Tomasz Krakowski · Rabia Tahir · Abdallah S. Daar

Received: 4 February 2011 / Revised: 27 April 2011 / Accepted: 17 June 2011

Abstract Post-Human Genome Project progress has enabled a new wave of population genetic research, and intensified controversy over the use of race/ethnicity in this work. At the same time, the development of methods for inferring genetic ancestry offers more empirical means of assigning group labels. Here, we provide a systematic analysis of the use of race/ethnicity and ancestry in current genetic research. We base our analysis on key published recommendations for the use and reporting of race/ethnicity which advise that researchers: explain why the terms/categories were used and how they were measured, carefully define them, and apply them consistently. We studied 170 population genetic research articles from high impact journals, published 2008–2009. A comparative perspective was obtained by aligning study metrics with similar research from articles published 2001–2004. Our analysis indicates a marked improvement in compliance with some of the recommendations/guidelines for the use of race/ethnicity over time, while showing that important shortfalls still remain: no article using 'race', 'ethnicity' or 'ancestry' defined or discussed the meaning of these concepts in context; a third of articles still do not provide a rationale for their use, with those using 'ancestry' being the least likely to do so. Further, no article discussed potential socio-ethical implications of the reported research. As such, there remains a clear imperative for highlighting the importance of consistent and comprehensive reporting on human populations to the genetics/genomics community globally, to generate explicit guidelines for the uses of ancestry and genetic ancestry, and importantly, to ensure that guidelines are followed.

Keywords Race · Ethnicity · Ancestry · Genetic ancestry inference · Terminology · Genetic research

S. E. Ali-Khan (✉) · T. Krakowski · R. Tahir · A. S. Daar
McLaughlin-Rotman Centre for Global Health, University Health Network and University of Toronto, 101 College St, Suite 406, Toronto, ON M5G 1L7, Canada
e-mail: sarah.ali-khan@mrcglobal.org; sarahalikhan@me.com

A. S. Daar
e-mail: abdallah.daar@mrcglobal.org

A. S. Daar
Department of Public Health Sciences and of Surgery, University of Toronto, Toronto, ON M5S 1A8, Canada

A. S. Daar
McLaughlin Centre for Molecular Medicine, University Health Network and University of Toronto, Toronto, ON M5S 1A1, Canada

A. S. Daar
Grand Challenges Canada, http://www.grandchallenges.ca

Introduction

The completion of the Human Genome Project over a decade ago has led to intensified studies of genetic variation in human populations. Much of this work uses specific population identities to categorize groups, for example Caucasian, Korean, South Asian and Yoruban, and in addition often uses the generic terminology 'race' and 'ethnicity' to refer to them. The validity of using socially-visible groups in biomedical research has been an ongoing controversy. However, there is now resurgent interest in the subject (Burchard et al. 2003; Cooper et al. 2003; Duster 2005; Risch et al. 2002; Schwartz 2001; Stevens 2003), because technology advances are increasing opportunities to clarify the relationship between social identity, genetic diversity and health, and to move beyond old prejudices

Published online: 07 July 2011

about human difference (Ali-Khan and Daar 2010; Editorial 2004a; Rotimi 2004). Accordingly, in the last 2 years several multidisciplinary groups including our own, have convened to examine these issues afresh (Caulfield et al. 2009, Lee et al. 2008).

A fundamental difficulty raised by the use of socially-visible population labels—whether they are referred to as 'races', ethnicities, nationalities, or by other language—is that their meanings and parameters are context-dependent (Kressin et al. 2003; Rotimi 2004), and have powerful ramifications beyond the domain of science (Bamshad and Olson 2003; Clayton 2002; Gould 1981; Lewontin 1995; Provine 1973). Lack of clarity and consistency in the description of research populations and inadequate justification for their use has been a persistent source of concern in biomedical research (Bhopal 1997; Clayton 2002; Collins 2004; Comstock et al. 2004; Editorial 2004b; Lee 2004; Sankar and Cho 2002), and can have adverse scientific and social consequences, particularly in the context of genetics research. As such, failure to define a group label or describe how membership was ascertained makes it difficult to know who exactly is being studied, challenging the reproducibility of research findings and limiting portability to other geneticists, disciplines, and the clinic (Brown 2007; Editorial 2004b; Sankar et al. 2007). Further, such ambiguity can encourage racial/ethnic stereotypes and over-simplifications that stymie, rather than promote, understanding of genetic diversity (Bamshad et al. 2004; Race Ethnicity and Genetics Working Group 2005). Likewise, failing to explain why a particular population was studied with respect to the research question can imply that social identity is the basis for any observed phenotypic differences. Such interpretations may divert from further study to identify true underlying mechanisms (Sankar et al. 2004), and have dangerous clinical consequences by encouraging reliance on social identity for prescription or prognosis (Braun et al. 2007; Geiger 2003; Lee 2005).

Ongoing concern has prompted journal editors, professional societies and expert commentators to repeatedly offer guidelines for the use and reporting of race and ethnicity in genetic research. These have largely converged on four key points; (1) define the race and ethnicity, or more broadly the population terms, used in the context of the study (Anonymous 2005; Burchard et al. 2003; Cooper et al. 2003; Editorial 2004b; Iverson et al. 1998; Kaplan and Bennett 2003; Race Ethnicity and Genetics Working Group 2005; Sankar and Cho 2002; Winker 2004); (2) explain how the terms or categories relate to the research hypothesis, or why the particular population was chosen for study by the researchers (Anonymous 2005; Editorial 1996, Editorial 2004a; International Council of Medical Journal Editors 2010; Iverson et al. 1998; Kaplan and Bennett 2003; Lee et al. 2008; Race Ethnicity and Genetics Working Group 2005; Rivara and Finberg 2001; Sankar and Cho 2002; Winker 2004); (3) describe how participants were assigned to the research populations (Anonymous 2005; Editorial 2004b; Lee et al. 2008; Race Ethnicity and Genetics Working Group 2005; Sankar and Cho 2002; Winker 2004); and (4) describe the limitations of the study with respect to the populations to which the research findings can be generalized (Anonymous 2003; Anonymous 2005; Davis et al. 2001; Ioannidis et al. 2004; Osborne and Feit 1992). Various of these have been endorsed by biomedical journals, and by the International Council of Medical Journal Editors http://www.icmje.org/journals.html#S (for review, see Caulfield et al. 2009). However, studies assessing compliance in genetic research published over 2001–2004 (Editorial 2004b; Sankar et al. 2007; Shanawani et al. 2006); indicated that guidelines were not widely followed.

Since those data were collected, the advent of high-resolution genome-wide genotyping is allowing more empirical description of individuals and populations, by the inference of genetic or 'biogeographical' ancestry (Bamshad et al. 2004; Li et al. 2008; Novembre et al. 2008; Rosenberg et al. 2002; Royal et al. 2010; Shriver et al. 2004; Via et al. 2009). Used to determine and quantify genetic background, this technology can augment or supersede the use of proxy methods, such as self-identified race/ethnicity, physical appearance, language-spoken, or ancestry based on geographical origin, to stratify research participants and maximize their relative genetic homogeneity. Thus, some have suggested the use of 'ancestry' rather than race/ethnicity to describe group differences and genetic variation, because of its more objective basis, and perceived distance from negative connotations associated with 'race' (Ali-Khan and Daar 2010; Bamshad et al. 2004; Race Ethnicity and Genetics Working Group 2005; Smart et al. 2006). Our study had two goals; (1) to assess current compliance with recommendations for the use of race and ethnicity—or more broadly social identity—in genetic research; and (2) to examine the use of 'ancestry' as a generic terminology to describe study populations, and also in the sense of 'genetic ancestry' by the use of empirical genomic methods to categorize research groupings.

Authors who previously examined the use of race and ethnicity in genetic research considered all specific population identifiers used in the context of humans as 'race and ethnicity' terms (Sankar et al. 2007; Shanawani et al. 2006). We do not disagree with this, but for the purposes of the second part of our analysis we went beyond previous analyses, sub-dividing our data by the generic terminology used to refer to the specific named study populations in articles, in order to compare articles which used the generic terms 'race and/or ethnicity', with those using 'ancestry', or 'other' terms. In addition, we note that we did not define

race or ethnicity or ancestry for the purposes of this work, but rather kept the study open-ended with the goal of observing how these terms are currently put to use by authors. In this vein, we add that the goal of this study was not to assess which of these terms should be used by authors. However, we agree with previous commentators that the study of DNA within the context of socially-identified groups in no way justifies the definition of sub-groups of individuals as biologically distinct races (Collins 2010).

Materials and methods

Study design

In this work we undertook a systematic analysis of scientific articles reporting genetic research in the context of human populations. Our analysis was divided into two parts. In the first part, to evaluate how use and reporting of this research has changed over time we assessed selected metrics adapted from previous studies (Sankar et al. 2007; Shanawani et al. 2006). We also asked new questions to examine the use of 'ancestry' to describe populations, the use of genotyping data to assign ancestry and thus verify research group membership, and whether discussion of social and ethical implications of the reported research was included in articles. In the second part of the study, to assess differences between articles using different generic terminology to refer to study populations we sub-divided the data by articles using; (1) *race and/or ethnicity*; (2) *ancestry*; and (3) *other terminology*. We then compared the metrics obtained in part one of the study across these sub-divisions of the data. We also collected qualitative data with respect to how 'ancestry' was used in articles.

Sample selection

We conducted a Pubmed search strategy to obtain a sample of journal articles for analysis. We used the keywords (race OR ethnicity OR ancestry) and the genetic terms (polymorphism OR CNV OR SNP), with the limits; humans, English, and publication dates between January 1st 2008 and December 31st 2009 ($N = 3536$). The use of 'race', 'ethnicity' and 'ancestry' in Pubmed captures articles in which these words occur in the text, articles that use these as words as MeSH headings, and the hierarchy of terms occurring under these headings. For example, 'race', is a synonym for the MeSH heading 'Continental Population Groups', which includes; 'American Continental Ancestry Group'; 'American Native Continental Ancestry Group'; 'Asian Continental Ancestry Group'; 'European 'Continental Ancestry Group'; and 'Oceanic Continental Ancestry Group'. Likewise, each of these terms captures all the specific groups classified to these geographic regions. For example 'American Continental Ancestry Group' includes; 'Indians Central America'; 'Indians North America', 'Indians South America'; and 'Inuits'; and likewise, when expanded each of these terms captures a range of specific population identifiers. For example 'Inuits' corresponds to; 'Inuit'; 'Inupiat(s)'; 'Eskimo(s)'; 'Kalaallit(s)'; and 'Aleut(s)' (see: http://www.ncbi.nlm.nih.gov/mesh).

We decided to direct our sample toward articles that are most likely to reflect the state of the art in the field of genetic research, and to be of high quality. Our rationale was that such articles may be most likely to have wider scientific and social influence, by serving as models and hypothesis generators for other researchers, and by infiltrating the non-geneticist community by being reported in the popular press. To capture articles most likely to be of this type, we identified a convenience sample of 10 leading population-based geneticists, genetic epidemiologists and genome scientists based in the United States and Canada, and asked them to rate the top 5 most influential journals in which to publish their work. We then limited our article collection to the top 6 highest ranked journals from this survey. These were; the American Journal of Human Genetics; Human Genetics; Nature; Nature Genetics; PLoS Genetics; and Science ($N = 197$) (search completed February 2010) (see Table 1). We note that three of these have published policy on the use and reporting of race and ethnicity (Brown 2007; Editorial 2004a, b). However, none are listed as explicitly endorsing the ICMJE's Uniform Requirements for manuscripts submitted to biomedical journals (see http://www.icmje.org/journals.html#S).

Table 1 Sample set characteristics, *N* (%)

Total sample	$N = 170$
Year of publication	
2008	$N = 93$ (54.7%)
2009	$N = 77$ (45.3%)
Journal of publication (2008 impact factor)	
American Journal of Human Genetics (10.153)	$N = 41$ (24.1%)
Human genetics (4.042)	$N = 38$ (22.4%)
Nature (31.434)	$N = 13$ (7.6%)
Nature Genetics (30.259)	$N = 40$ (23.5%)
PLoS Genetics (8.883)	$N = 32$ (18.8%)
Science (28.103)	$N = 6$ (3.5%)
Article general field of interest	
Population genetics	$N = 26$ (15.2%)
Medical	$N = 127$(74.7%)
Methods	$N = 9$ (5.3%)
Non-medical	$N = 8$ (4.7%)

The abstracts and MeSH information for each article were then reviewed to exclude all but original research articles from the sample—news, comments, letters, reviews and meta-analyses were removed. Finally, the entire articles were downloaded and reviewed in detail to verify each described original research studying human genetic variation using human tissue samples or human subjects. This process yielded a final study sample of 170 articles for analysis.

Data analysis

Part one

The study articles were saved as PDFs, and printed out, read and examined by hand to extract data. In addition, the full Medline format information on each article was uploaded to a Refworks database, and the data from our analysis were recorded in customized fields. To enable comparison to previous study on articles published from 2001 to 2004 (Sankar et al. 2007, Shanawani et al. 2006), the coding and analytical framework we used was adapted principally from Sankar et al. (2007), and with reference to the analysis by Shanawani et al. (2006). Sankar's group developed content codes to analyze how the research populations were described, and the main components and structure of scientific articles. In addition to using these, we developed additional codes to assess the use of ancestry, of empirical genomic methods to measure ancestry, or assign or verify membership in research populations, and the discussion of ethical and social aspects in articles. An initial set of codes was tested by SEA, RT and TK. These codes were subjected to several rounds of consensus coding (Jenkins et al. 2005; Sankar et al. 2007) and discussion amongst all the authors. When interpretation and conceptual issues were resolved, and the codes were deemed to adequately capture relevant article features, a coding guide was generated listing coding rules, definitions and examples. The final study analysis was carried out by SEA.

Coding

The analysis codes evaluated four main areas: (1) basic article features; (2) reasons researchers gave for how and why they used named populations in the study design; (3) the role of the named populations in the research design or the description of the research; (4) use of empirical genomic means to assign or verify membership in the research populations; and (5) discussion of social or ethical implications of human genetic research. We analyzed each article by looking for text corresponding to these codes, or pieces of information, as described below and scored them as a yes/no variable. Additionally, for many of the codes, we collected qualitative data for further analysis, by recording the text content as well. We also noted the country of the institution of the first author, how the research was funded, whether or not informed consent was reported for the research populations involved, and whether or not a conflict of interest statement was provided.

Basic features Each article was analysed with respect to three basic features providing fundamental information about the study it reported. Each code was scored as a yes/no variable. (1) *hypothesis* was defined as the presence of a founding idea or assumption stated as the starting point for investigation. Text identified for this code included formally stated hypotheses, and more general research questions, goals or aims. In each case the text had to state or imply that the idea provided the basis for the study; (2) *limitations* were statements that described the factors that restricted the generalizability of study findings. Statements had to be explicit and related to study design to be coded as limitations. Hypothesis and limitations are standard aspects of scientific research articles. Inclusion of a specific hypothesis is important as this is where readers might expect to find an explanation for how identifying a study population as a specific race, ethnicity or ancestry group relates to the study premise or research question. A limitations statement offers the opportunity to explain how widely the findings can be applied to populations beyond the study sample that might be associated with the race, ethnicity or ancestry terms used in the study. Note that some articles were not included in the limitations analysis because we judged their analysis to not require such a qualification; and (3) *sample origin* was defined not as the geographical region from which the samples were obtained, but where and how the researchers acquired the tissue samples or genetic data, for example—whether they were obtained from a tissue or databank, collected at a hospital, or were already in researchers' possession.

Reason for using populations To examine authors' explanations for why research was conducted using race and ethnicity or ancestry terms, articles were classified based on three features that have been recommended by expert commentary, journals, and professional societies (Ali-Khan and Daar 2010; American Academy of Pediatrics: Committee on Pediatric Research 2000; American Anthropological Association 2000; Bamshad et al. 2004; Editorial 1996; International Council of Medical Journal Editors 2010; Lee et al. 2008; Race Ethnicity and Genetics Working Group 2005; Rivara and Finberg 2001; Winker 2004), and see, http://www.icmje.org/urm_full.pdf. (1) *Why populations* was used to label text that gave reasons for pursuing the research question by using a population so identified; (2) *Why this population* was used

to code reasons provided for studying the particular population(s) in question. Reasons could be practical (e.g. because the sample was available) or theoretical (e.g. because the condition of interest was known to occur frequently in a particular group); (3) *Basis for assigning population term* was defined as the method by which membership in the study population was determined, or the population label was assigned to research participants. For example, self-reported by subjects, taken from existing records, assumed because of the geographical region where subjects were recruited, or assigned based on genomic inference. If an article provided any of these means it was coded yes. Thus, a yes/no variable and how, as qualitative data, was collected.

Use of genotyping data to infer genetic ancestry To begin to evaluate the nature and the degree to which high resolution genome-wide genotyping—or genetic ancestry testing—is being used to assess the genetic background or ancestry of research participants or samples, we labelled text that described such methodologies. Only studies using these approaches as part of their process of assigning or verifying the membership of participants or samples to research groupings, or to assess for population stratification were coded as 'yes'. The use of such methods to analyse the genetic structure of populations as the main goal of the reported research were coded 'no'. Both a yes/no variable and how, as qualitative data, was collected.

Defines race and/or ethnicity, or ancestry terms To assess the degree to which authors defined and described the terms and identities used to refer to research populations we labelled text according to the following codes (1) *Defines generic* '*race*', '*ethnicity*' *or* '*ancestry*' was applied when text explicitly defined race, ethnicity or ancestry as a genetic, social or biological concept, or provided a reference to information that did so. Further, we began to assess the comprehensiveness of research population definitions. We based our criteria on those recommended by commentators who underlined the importance of thorough, and multi-dimensional description of populations (Bamshad et al. 2004, Editorial 2004b). Thus, we applied (2) *Defines specific race or ethnicity, ancestry term (or population identifier)* to articles only when they provided all of the following information; (1) the identifier or name of the research population; (2) the geographical location where the participants were recruited or the community where they were resident; (3) their 'racial', ethnic, or geographical ancestral origin; and (4) specified *how* this label was assigned (for example, by self-report, by genomic ancestry inference, based on multiple generations of the participants' family etc.). In addition, if text, or a figure (e.g. a principal components plot) described the group genomically, defining parameters for group exclusion or inclusion, this was also coded as yes.

The role of named populations in genetic research To examine the various ways that articles used race and ethnicity, or ancestry, text was labelled that referred to the following 5 codes: (1) *Label for study population only* was applied to text where race, ethnicity, ancestry or other populations terms were used to label the study population only, and not as a research variable; (2) *Independent* and; (3) *Dependent* were applied respectively when race, ethnicity, ancestry or other population terms were employed as independent or dependent variables in the research being reported; (4) *DNA with label* indicated where authors had labelled DNA—for example, alleles, chromosomes, haplotypes, or mutations—with a race, ethnicity or ancestry term, as in 'Mexican and Caucasian T allele (Plaisier et al. 2009) Codes (1–4) could co-occur, but codes (2) (Independent) and (3) Dependent) were mutually exclusive.

Social and ethical implications related to human population genetic research We looked for statements discussing social or ethical implications of population-based genetic research. Such content had to discuss implications arising from the genetic research being reported—e.g. text relating to the potential for study results to stigmatize the research population. We coded these as a yes/no variable, and if found, the issues discussed were recorded as qualitative data.

Categorization of articles by general field of interest We also categorized the articles by their general field of interest; 'population genetics' was defined as including population genetic and studies examining inter or intra-population genetic structure, genetic anthropology, and whole genome sequencing articles; 'medical' included disease and pharmacogenomics-related articles; 'methods', reported new methodologies or analytical approaches in genetic research; and 'non-medical' was defined as articles reporting studies of non-medical-related phenotypes, for example height or hair colour. For the purposes of this analysis these categories were exclusive.

Part two

Generic terminology used to refer to research populations We recorded all the generic terminology used to describe the research populations in each study, and the specific research population names or identifiers. If an article referred to the research populations by 'race' or 'ethnicity/ethnic' anywhere in the main article body or supplementary materials we coded the article as *'race and/or ethnicity'*. Likewise, if an article referred to populations

as ancestries or ancestry groups (but not race or ethnicity), it was coded *'ancestry'*. Articles using only a specific population identifier such as whites, African Americans, Han Chinese etc., or that described populations using any other terminology such as origin, descent, etc. were coded *'other'*. We also recorded examples where the generic terms 'race and ethnicity' were used synonymously with 'ancestry', or that used them in a conceptually distinct fashion, and collected qualitative data regarding the use and application of 'ancestry' in articles.

We note that for the purposes of this analysis and consistent with others (Sankar et al. 2007; Shanawani et al. 2006), we did not distinguish conceptually between race and ethnicity. Despite some commentators offering distinct definitions of these (Editorial 2004b; Harrison 1995; Kalow 2001; Wood 2001), it appears that in practice they are most often used interchangeably (Condit 2007; Oppenheimer 2001; Sankar and Cho 2002). Thus, in this work, we considered them together as one category.

Supplementary and additional data

Many articles provided additional supplementary information or methods online. These were downloaded, examined for relevant information, and coded as part of the analysis for each article. Some referred readers to previously published literature for details about research procedures or the study population. These articles were also downloaded, examined and relevant statements were used as the basis for assigning the codes to the original article. If these articles did not provide the relevant details but in turn referenced another paper, we scored the article as 'no' for the code in question.

Statistical analyses

After sub-dividing the data by the generic terminology used to refer to the study populations in articles, as in (1) *race and/or ethnicity*; (2) *ancestry*; and (3) *other*. We then compared frequencies of individual codes across the resulting subsets of the data, assessing the significance of any differences via the chi-square statistic. Statistical tests were performed using SigmaStat statistical software (Version 3.5).

Results

Sample set characteristics

We reviewed and analyzed 170 research articles published in 2008 and 2009 reporting genetic research in the context of human groups. Basic characteristics and categorization by the articles' general field of interest are shown in Table 1.

Part one—compliance with recommendations for the use and reporting of populations in genetic research

Basic article features

We were able to identify a clearly stated hypothesis or research questions in almost every article in our sample (99.4%, $N = 169$) (Table 2). Likewise, most papers described the origin of their research samples. Fewer articles described the limitations of their studies with respect to the populations investigated (52.4%, $N = 87$). Most of these limitations statements were not extensive, but rather comprised of a sentence in the article discussion stating that the study findings should be validated or further investigated in diverse populations, or in other 'racial', ethnic or ancestry groups (see for example, Ganesh et al. 2009).

Reason for using populations

About two thirds of articles explained why they chose to study labelled populations (65.9%, $N = 112$), or why they chose to study the particular populations featured in the research (68.8%, $N = 117$) (Table 2). Most of these explanations were based on the phenotype or condition under study being of high prevalence in the study population, or the fact that this group was understudied in comparison to others, for example 'Because neuroblastoma in the United States is demographically a disease of Caucasians of European descent, we limited our initial analyses to this racial group to minimize phenotypic variability'(Diskin et al. 2009). A key scientific consideration in selecting samples for association studies is that they be drawn from the most genetically homogeneous population possible—thus striving to avoid spurious associations resulting from population stratification (Cardon and Palmer 2003; Marchini et al. 2004). However, few articles (4.7%, $N = 9$), specifically linked this notion to the use of labelled populations in their study, or to the particular population investigated. Of the articles that did not explain why they chose to study labelled populations, all but one were association studies or other analyses to identify a trait's genetic basis. In many of these articles populations/samples were ostensibly used because of their availability to researchers, although this was not explicitly stated.

Most articles also provided some basis for how the population label was assigned to research participants (88.2%, $N = 150$). Most indicated that this was by self-reported race, ethnicity, geographical origin or ancestry, and/or was determined based on the geographical region where participants were recruited or resided, and/or was assigned or verified using genomic data (see following section). Assigning population labels on the basis of more than one

Table 2 Sample set coding frequencies, *N* (%)

Variables coded	*N* = 170 (%)
Basic features	
Hypothesis	169 (99.4%)
Limitations	87 (52.4%)
Sample origin	163 (95.9%)
Reason for using population	
Why populations	112 (65.9%)
Why this population	117 (68.8%)
Basis for assigning population label	150 (88.2%)
Use of genotyping data to infer genetic ancestry	88 (51.8%)
SNP genotypes or ancestry informative markers (AIMs) used to infer ancestry proportions of individual participants' DNA samples	20 (23.3%)
Genotype data used to assess the genetic homogeneity of population by principal components cluster analysis, Samples outlying from population clusters of interest excluded from further analysis	36 (41.9%)
Text briefly states that potential population stratification was examined in the research populations, but no further details are provided	32 (36.4%)
Defines generic 'race and ethnicity' or 'ancestry'	0 (0%)
Defines specific population label/describes population group	102 (60.0%)
Ways of using populations in research	
Label for study population only	78 (45.9%)
Independent variable	87 (51.2%)
Dependent variable	1 (0.59%)
DNA with a label	23 (13.5%)
Discusses social and ethical implications	0 (0%)

generation of the research participants' family has been recommended (Tang et al. 2005). However, only 18 articles (10.6%) in our sample described using such approaches.

Use of genotyping data to infer genetic ancestry

Just over half the articles (51.8%, $N = 88$) described using genomic data to assess the genetic ancestry of research participants to assign or verify the research groupings, and/or to guard against population stratification (Cardon and Palmer 2003; Marchini et al. 2004) (Table 2). This is important because such approaches can substantiate the genetic similarity of individuals stratified using proxy methods, and provides another element to the description of research populations.

Most of these determinations were described in the methods section of articles where statistical analyses or quality control issues were described, and fell into three broad categories; (1) genome-wide SNP genotypes or ancestry informative markers (AIMs) were used to infer the ancestry proportions of individual participants' DNA samples. Those whose ancestry percentages fell below a specified cut-off were excluded from further analysis (23.3%, $N = 20$), see for example, (Trevino et al. 2009); (2) genome-wide SNP data was used to assess the genetic homogeneity of study populations, by principal components cluster analysis, sometimes in comparison to HapMap reference populations. Samples outlying from population clusters of interest were excluded from further analysis (41.9%, $N = 36$), see for example, (Yamaguchi-Kabata et al. 2008); (3) text briefly states that potential population stratification was examined in the research populations, but no further details are provided. These articles simply state that population genetic structure was not evident, or that it was found and corrected (36.4%, $N = 32$). Articles featuring this latter wording (3), were not coded 'yes' as providing the basis for assigning the population label, because no details were provided as to how samples were included or excluded from the research groups. Likewise, such text was not coded 'yes', as constituting a genomic description of the population for the same reason (see section below). Thus, genetic ancestry testing was described in a variety of ways, and at varying levels of detail by authors.

Defining race or ethnicity, and ancestry

No article in our sample set specifically defined the meaning of the generic terms 'race', 'ethnicity' or 'ancestry' in the research reported (Table 2). This was surprising to us given the high percentage of articles which

explained the basis for how they assigned the population label used (88.2%). Further, we expected that articles using terms of race, ethnicity or ancestry to categorize their research samples, or that used groups so labelled as independent research variables (51.2%, see section below), would also discuss or define the meaning of these concepts in the context of their study. Although no article provided explicit definitions, several studies whose goal was to analyze genetic substructure in populations, did begin to outline a distinction between population identifiers/names or self-identified ethnicity, and ancestry, which was framed in terms of genetic background (see for example, Li et al. 2008; Reich et al. 2009; Tishkoff et al. 2009).

A critical component of many recommendations has been to use as specific population labels as possible, to carefully define their meanings (Bamshad et al. 2004; Kaplan and Bennett 2003; Sankar and Cho 2002), providing as much information on the population 'as is compatible with ethical review board requirements' (Editorial 2004b). For the purpose of this study, we considered a definition to include (1) the name or population identity of the group; (2) the geographical region of recruitment or the community in which the research participant resides; (3) their ethnic identity and or/the geographical origin of their ancestors; and (3) a specific indication of *how* the latter was determined. To be scored yes, an article needed to provide all of this information. More than half the articles in our dataset defined the specific population identifier used according to these parameters (60%, $N = 102$). Of these, 54.9% ($N = 56$) included a genomic description (i.e. groups (1) and (2) described in the previous section). Of articles that did not 'define' the population, many noted the geographical location of recruitment or residence, and/or the race/ethnicity, or ancestry of participants or samples, but not *how* these latter categorizations were determined. For example, they might state that 'all subjects were of full Japanese ancestry' (Yasuda et al. 2008), but not explain precisely what this meant in context, or how it was determined.

Ways of using labelled populations in genetic research

About half of the articles (51.7%, $N = 88$) used the named populations as either dependent or independent variables (Table 2). The remainder used population identifiers only to label their study populations, but not to test a hypothesis related to the named group. A number of papers (13.5%, $N = 23$) labelled DNA by population. In most of these cases the population identifier was used to label the inferred ancestral identity of DNA sections in admixture mapping or similar studies, see for example (Hancock et al. 2009). A few articles used wording such as 'ethnic-specific locus' (Lei et al. 2009) or an 'Asian mitochondrial DNA haplotype' (Keyser et al. 2009).

Social and ethical implications

No articles mentioned or discussed social or ethical implications arising from genetic research in general, or from the research being reported (Table 2). On the one hand this was not really surprising given that geneticists and social scientists have not traditionally collaborated, despite calls for interdisciplinary perspectives (Ali-Khan and Daar 2010; Bonham et al. 2005; Condit 2007; Lee et al. 2008; Via et al. 2009). Conversely, given that authors from both disciplines have engaged these issues (Bamshad et al. 2004; Burchard et al. 2003; Caulfield et al. 2009; Cooper et al. 2003; Duster 2005; Lee et al. 2008; Rotimi 2004), we anticipated finding some discussion, however cursory, in articles.

Comparison of current data with previous studies

To begin to get perspective on how researchers' reporting has changed over the course of the past decade, we compared our findings with a previous study which analyzed genetic research articles published over 2001–2004, and with which we specifically aligned part of our study methodology and analysis (Table 3). We found marked increases in the numbers of articles providing hypothesis statements for the research reported (30% in the earlier study, compared to 99.4% of articles in our dataset, $P = <0.001$), and likewise describing the origin of their research samples (62.4% compared to 95.9%, $P = <0.001$). Compared to earlier in the decade, more authors described the limitations of their research findings with respect to the population-based data reported (22.7% compared to 52.4%, $P = <0.001$). However, still only about half the articles provided limitations.

There were substantial increases in the proportion of articles (1) explaining why samples/participants in the study were grouped by race and ethnicity, or ancestry labels (10.9% compared to 66.5%, $P = <0.001$); and (2) justifying why these particular population groups were studied (11.2% compared to 68.2%, $P = <0.001$). In contrast, there was no change over time in articles defining the generic terms 'race', 'ethnicity', or 'ancestry' in the context of their study—no articles in either dataset defined these terms. This was surprising given the intensification of population studies using these terms in the last 10 years, and the continued scrutiny of measurement, communication and identity issues over this time (Caulfield et al. 2009; Clayton 2002; Foster and Sharp 2004; Lee et al. 2008; Rotimi 2004). It also suggests that researchers consider the meanings of these terms self-evident.

Part two: cross-comparison of articles using different terminologies—race and ethnicity, ancestry or other

We recorded the generic terminology used to refer to research populations in each article. Most described the

Table 3 Comparison of current data with earlier study

	Sankar et al. (2007)	Current study	
Data derived from articles from publication years	2001–2004	2008–2009	
# Articles	330	170	
Sample selection criteria	Medline search strategy: race and ethnicity, genetics and population keywords; AND publication in one of 3 journal type samples (genetics, clinical, and general); mainly high impact journals	Pubmed search strategy: (race OR ethnicity OR ancestry) AND (SNP OR polymorphism OR CNV) keywords; AND publication in one of six leading journals for the publication of human genetic research; mainly high impact journals	
Variables coded			
Basic features			*P* value by chi sq
Hypothesis	99 (30%)	169 (99.4%)	<0.001*
Limitations	75 (22.70%)	87 (52.4%)	<0.001*
Sample origin	206 (62.40%)	163 (95.9%)	<0.001*
Reason for using populations			
Why populations	36 (10.90%)	113 (66.5%)	<0.001*
Why this population	37 (11.20%)	116 (68.2%)	<0.001*
Defines generic 'race and ethnicity' or 'ancestry'	0%	0%	N/A
Ways of using populations in research			
Label for study population only	76 (23%)	82 (48.2%)	<0.001*
Independent variable	154 (46.70%)	87 (51.2%)	0.389
Dependent variable	16 (4.80%)	1 (0.59%)	0.026*
DNA with a label	35 (10.60%)	23 (13.5%)	0.412

* Indicates statistically significant difference, $P < 0.05$

studied populations as races or ethnicities ($N = 80$, 47.1%) (Table 4). For example, 'The 67 populations analyzed in this study represent 41 ethnic nationalities living in China and other eastern Asian regions' (Shi et al. 2009), and 'All genetic association analyses were stratified by self-identified race (white vs. African American)…'(Rasmussen-Torvik et al. 2009). Only 4 articles (2.4%) used race only as a terminology, the rest used 'race/ethnicity', or ethnicity only, to describe populations. 'Ancestry' or 'ancestry groups' were used in 22.4% of articles ($N = 38$), for example, 'Risk allele frequencies of rs12970134 are higher among individuals of Indian Asian ancestry than those of European ancestry' (Chambers et al. 2008). The remainder of articles referred to populations only by a specific population identifier or name such as 'European American', 'Hadza' or 'Japanese', or by using various descriptors—most often origin, descent and derived, for example 'this difference is maintained in American children of Japanese descent resident in the US' (Burgner et al. 2009) (see Table 5 for a list of terms and ways of describing populations compiled from our sample set). We cannot comment on how the relative use of these terminologies has changed over time, as previous studies did not examine this parameter.

To assess potential differences in the way researchers use 'race and/or ethnicity', compared to 'ancestry' or 'other' kinds of terminology to describe research populations or samples, we sub-divided our data by the generic terminology used, and analyzed the frequency of our research codes and categorizations across these sub-groups (Table 4). For the basic article features, there was no significant difference between terminology sub-groups. However, articles using race and/or ethnicity were significantly more likely to provide a justification for why the research studied populations so labeled (75.0%, $N = 60$) compared to those using ancestry (44.7%, $N = 17$) or other terminology (69.2%, $N = 36$) ($P = 0.004$). Articles using race and/or ethnicity were also significantly more likely to report medical-related research ($P = <0.001$). Consistent with these findings, during our analysis we noted that medical–related articles often investigated health disparities between groups framed in terms of race or ethnicity, and rationalized the study of their research populations on this basis. On the other hand, articles using ancestry were less likely to provide a justification for the use of this terminology to label populations, or why particular populations were studied (Table 4).

Table 4 Presence of coded article features by generic terminology used

Total sample set, $N = 170$	Race and ethnicity $N = 80$ (47.1%)	Ancestry $N = 38$ (22.4%)	Other $N = 52$ (30.6%)	P value by chi sq
Variables coded				
Basic features				
Hypothesis	79 (98.8%)	38 (100%)	52 (100%)	0.568
Limitations	43 (55.1%)	16 (45.7%)	28 (53.8%)	0.319
Sample origin	75 (93.8%)	37 (97.4%)	51 (98.1%)	0.413
Reason for using population				
Why populations	60 (75.0%)	17 (44.7%)	36 (69.2%)	0.004*
Why this population	55 (68.8%)	23 (60.5%)	38 (73.1%)	0.372
Basis for assigning population label	71 (88.8%)	35 (92.1%)	44 (84.6%)	0.542
Use of empirical genomic methods	44 (55.0%)	25 (65.8%)	19 (36.5%)	0.017*
Defines generic 'race and ethnicity' or 'ancestry'	0 (0%)	0 (0%)	0 (0%)	
Defines specific population label	51 (63.8%)	24 (63.2%)	27 (51.9%)	0.361
Ways of using populations in research				
Label for study population only	34 (43.0%)	21 (56.8%)	27 (51.9%)	0.352
Independent variable	46 (58.2%)	17 (45.9%)	24 (46.2%)	0.296
Dependent variable	0 (0%)	0 (0%)	1 (1.9%)	0.319
DNA with a label	14 (17.5%)	3 (7.9%)	6 (11.5%)	0.319
General article field of interest				
Population genetics	10 (12.5%)	1 (2.6%)	15 (28.8%)	0.002*
Medical	68 (85%)	32 (84.2%)	27 (51.9%)	<0.001*
Methods	1 (1.3%)	2 (5.3%)	6 (11.5%)	0.036*
Non-medical	1 (1.3%)	3 (7.9%)	4 (7.7%)	0.134
P value by chi sq for terminology used within field of interest	<0.001*	<0.001*	<0.001*	

* Indicates statistically significant difference, $P < 0.05$

Finally, we hypothesized that articles using 'ancestry' to label populations would be more likely to use genotyping data to assess the genetic background of their research groups, in order to assign the population label. Indeed, there was a significant relationship between the use of 'ancestry' and of empirical genomic methods (65.8%), compared to 'race and ethnicity' (55.0%), or articles using 'other' ways of referring to their research populations or samples (36.5%) ($P = 0.017$) (Table 4). The importance of controlling for population stratification through the assessment of genetic ancestry has been a key consideration in the context of genetic association studies (Cardon and Palmer 2003; Marchini et al. 2004). Consistent with this, medical (53.3%) and non-medical-related articles (62.5%) in our sample set—of which 41.5% and 50.0%, respectively reported genome-wide association studies—were more likely to use genomic methods, while population-related articles were less likely to (23.1%) ($P = 0.006$) (Table 6). Again this was consistent with our observations that population genetics type articles—which often mapped genetic substructure across populations—mostly relied on language-spoken, geographical location of residence or self-identified ethnicity to assign group membership. Conversely, case–control studies aiming to identify new genetic variants and striving to minimize population stratification, most often analyzed genotyping data or inferred genetic ancestry to stratify samples.

Uses of ancestry in our sample set

To further understand how 'ancestry' is being by employed in research practice, we catalogued how the term was used by authors in articles. We found the terms 'ancestry' or 'ancestry group' were used in three main ways. Most commonly, they were used, as described above, to refer to the geographical origin of populations, for example 'individuals of European ancestry', or the line of heritage or descent of a group, for example, 'Ashkenazi Jewish ancestry' (Bronstein et al. 2008). In particular, ancestry was often used to describe populations/individuals for whom the geographic origin of their predecessors is different from their current place of residence (for example African Americans, or European Americans). 'Ancestry', 'geographic ancestry' or 'biogeographic ancestry' was also

Table 5 Terms used, and ways of describing populations compiled from our sample set

Terms and ways of describing or referring to populations	Example
Ancestry/ancestral groups	'Despite wide variation in allele frequency, these genetic variants show notable homogeneity of effect across populations of European ancestry living at different latitudes and show independent association to disease risk' (Bishop et al. 2009)
Anthropological names	'The names we use are the ones by which the groups are described anthropologically, but are not unique identifiers' (Reich et al. 2009)
'X'-derived	'Variants in the FTO gene have been associated with obesity measures in mainly European-derived populations' (Wing et al. 2009)
Of 'X'-descent	'Significant associations with individual SNPs at a common locus were observed in the two independent populations of African descent' (Garner et al. 2008)
Ethnicity/ethnic	'Importantly, we made similar observations when comparing populations of the same ethnicity' (Shi et al. 2009)
Ethnogeographic groups	'These results also show that two individuals carrying the same mtDNA haplotype can be classified in opposite ethnogeographic groups…' (Keyser et al. 2009)
Linguistic groups	'The structure results, population phylogenies, and PCA results all show that populations from the same linguistic group tend to cluster together' (HUGO Pan-Asian SNP Consortium et al. 2009)
Of 'X'-origin	'The clinical characteristics of participants in five independent cohorts—the white U.S. GWAS sample (n 1/4 1000), the white US family sample (n 1/4 1972), the Chinese hip fracture (HF) sample (n 1/4 700), the Chinese BMD sample (n 1/4 2995), and the Tobago cohort of African origin (n 1/4 908 men)—are described in Tables 1, 2, 3, 4, 5' (Xiong et al. 2009)
Race/racial groups	'We also performed race stratified analyses to control for potential confounding by race as well as to evaluate the previously reported race-specific results' (Crosslin et al. 2009)
Only population identifier or name used	'Using genome-wide association data from 1,376 French individuals, we identified 16,360 SNPs nominally associated with T2D and studied these SNPs in an independent sample of 4,977 French individuals' (Rung et al. 2009)

Table 6 Use of empirical genomic methods by article field of interest

General article field of interest	Population genetics ($N = 26$)	Medical ($N = 127$)	Methods ($N = 9$)	Non-medical ($N = 8$)	Chi sq
Used genomic methods	6 (23.1%)	74 (53.3%)	3 (33.3%)	5 (62.5%)	0.006*

used to refer to the genetic background of individuals, or to sections of DNA along a chromosome, as inferred by the analysis of multi-locus genotypes. Sometimes these latter applications were distinguished by being specified as 'genetic ancestry' (for example, see (Li et al. 2008)). Most often ancestry or genetic ancestry used in this sense was framed in terms of continental origin—African, European, Native American or Asian, as determined by the use of HapMap I populations or other continental reference SNP collections. However, a few studies analyzed the genetic ancestry of populations on a regional scale (see for example, (Novembre et al. 2008; Reich et al. 2009)).

Confusing or interchangeable uses of race and ethnicity and ancestry

Careful definition and precise use of terms used to refer to populations would facilitate clarity about who is being studied, aid in dissecting genetic from environmental influences on phenotype, and assist in deconstructing conflation between social identity, and genetic background. However, as noted, none of the articles in our sample set defined 'race and/or ethnicity' or 'ancestry' in the context of their reported research (Table 2). A minority of articles used both race/ethnicity, and ancestry, to refer to the same populations in their reported research (21.2%, $N = 36$). Of these, about half used the terms distinctly, for example, Choudhry et al. (2008) specified that the ethnicity of participants was Puerto Rican (based on the reported ethnicity of the participants' biological parents and all four biological grandparents), and then analysed their genetic ancestry in terms of West African, European and Native American background. However, some articles used race and ethnicity, and ancestry, interchangeably or indistinctly (Table 7). Most notably, while most articles which analysed genotypes to infer population genetic identities framed these in terms of ancestry, for example, 'genetically-inferred individuals of European ancestry (Trevino et al. 2009), a few articles described these in terms of race (Yeager et al. 2009), or ethnicity, (Glessner et al. 2009)

Table 7 Examples of indistinct, interchangeable or confusing usage of race and ethnicity and ancestry compiled from our sample set

Example from text	Comment
(1) 'To minimize confounding by ethnic variation we restricted our study population to individuals of self-reported European descent' (Amos et al. 2008)	Authors do not explain why or how 'ethnic variation' would confound results. The relationship between 'ethnic variation', 'self-reported European descent' and genetic background is not explicated. No term defined
(2) 'All genetic association analyses were stratified by self- identified race (white vs. African American) to avoid spurious associations due to population stratification' (Rasmussen-Torvik et al. 2009)	The relationship between 'self-identified race', and population stratification is not explicated. No term defined
(3) Research populations—Gullah, African American and European American—are referred to as being of African and European descent respectively in main article body, while in the supplementary text they are referred to as 'races' (Nath et al. 2008).	Use of differing terminology to refer to the same populations. 'Race' is not defined
(4) 'The self-identified race/ethnicity information for these AGRE individuals is listed below'; however, the table is entitled 'AGRE self-identified ancestry' and lists' American Indian/Alaskan Native; Asian; Black or African American; More Than One Race; Native Hawaiian or other Pacific Islander; Unknown; and White (Wang et al. 2009)	Interchangeable use of ancestry, and race and ethnicity. No term defined
(5) 'All samples must have Caucasian ethnicity based on hierarchical clustering of AIMs genotypes, and all other samples were excluded'. 'Ancestry' only, used in main article body, ethnicity used only in supplementary text (Glessner et al. 2009)	Authors are referring to the inference of population ancestral identity using empirical genomic methods. However, how 'ethnicity' relates to genetic background is not explicated. Inappropriate use of 'ethnicity', rather than ancestry. Use of anachronistic 'Caucasian', rather than 'European' terminology. No term defined
(6) 'Only subjects that self-reported as being of European ancestry were retained, regardless of their self-reported race'; Genetically inferred population identity referred to as 'imputed race' (Yeager et al. 2009)	Relationship between 'self-reported ancestry', 'self-reported race', and 'imputed race' not explicated. Inappropriate use of 'race' with respect to 'imputed race'. No term defined
(7) 'Distributions of racial ancestries were the same in cases and controls' (Walsh et al. 2008)	Inappropriate use of 'racial' and 'ancestry' together. No term defined

(Table 7). Finally, in a number of articles, ancestry or 'other' terms were used to refer to research populations in the main article body, while in supplementary materials race or ethnicity was used to describe the same populations. In some cases, this was because the research population's inferred genetic ancestry only was discussed in the article body, and the method—including the 'racial' or ethnic groups from which the 'ancestry' group was imputed—was provided in supplementary materials. In other cases, race, ethnicity and ancestry were used in indistinct and interchangeably ways in supplementary text suggesting that less care was taken in the preparation of these materials (Table 7).

Discussion

Recent advances in high resolution genetic analyses and access to larger and more diverse population samples are now offering unprecedented opportunities for biomedical progress, and for understanding human identities, histories and relationships. To maximize the benefits of this research, it is crucial that authors precisely define and describe who is under study, the constructs by which they are grouped, and how this is relevant to the research hypothesis. To evaluate the current state of research practice, we examined published articles with two goals: (1) to investigate how recommendations for the use of race and ethnicity—or more broadly social identity—in human genetic research are currently being followed; and (2) to examine the use of 'ancestry' as a generic terminology to describe study populations, and also in the sense of 'genetic ancestry' by the analysis of genomic data to stratify participants/samples.

We show that there has been marked improvement in compliance with many of the key published recommendations for the use and reporting of population-based genetic research over the last decade—at least in this sample of mainly high impact journals. However, our analysis highlighted considerable shortcomings. Below we discuss some of the main findings, and offer recommendations to improve on the current situation derived from our analysis (see Box 1).

'Ancestry' was used to refer to research populations in more than a fifth of articles in our sample set. More than 50% of articles used genetic ancestry inferences to assign participants/samples to research population groupings—most often the label 'ancestry' or 'genetic ancestry' was

Box 1 Recommendations for the genetics community and biomedical journal editors from our analysis, for the reporting of genetic research in human populations

(1) Provide a comprehensive explanation of the methods used for genetic ancestry imputations, including assumptions made, algorithms and parameters used, descriptions of population samples involved, and the limitations of inferences

(2) Define and differentiate the concepts of race, ethnicity, and ancestry used in the context of the reported research

(3) When empirical methods are used to assign ancestry labels, specify 'genetic ancestry' or 'inferred genetic ancestry' is being referred to, rather than simply 'ancestry'

(4) Provide an acknowledgment or brief discussion of social, ethical, legal, economic etc. issues raised by the reported research, if applicable

(5) Form a working group consisting of representatives from the spectrum of countries and cultures to engage the genetics community globally to:

- Highlight the importance of careful and consistent reporting on, and naming and description of, human populations in genetic research
- Address concerns and ambiguities in the implementation and reporting of genetic research in human populations
- Revise extant guidelines and explicitly generate guidelines for the uses of ancestry and genetic ancestry
- Gain broad endorsement of these guidelines/standards/requirements throughout the genetics community

(6) Ensure biomedical journals consistently enforce these standards and requirements in genetic research reporting

conferred in the context of such a genomic analysis. Genetic ancestry inferences were carried out in a variety of ways, and described with a variety of levels of detail. However, about a third of the articles in which genetic ancestry or population stratification was assessed and corrected, did not describe the method by which this was done at all (Box 1. (1) and (3)).

No article explicitly defined the meaning of the generic terms race, ethnicity or ancestry in context, or in relation to one another, even when both concepts were used within the same article. This was despite the terminology being used to label independent research variables in more than 50% of articles, and the acknowledged ambiguity of the construct of 'race' (Anonymous 2002; Long and Kittles 2009). Likewise, the concept of ancestry, despite its ostensibly objective basis, can be understood in multiple ways—for example genetic ancestry, geographical ancestry, biogeographical ancestry etc. (Royal et al. 2010; Via et al. 2009). Similarly, only one article explicitly discussed the relative and heuristic nature of inferred genetic ancestries and population models (Reich et al. 2009). Requiring authors to specifically define and differentiate of concepts of race, ethnicity, and ancestry would promote clarity for the reader about juxtapositions between genetic variation, population history and social identity. Equally productively, it might also engage researchers themselves in deeper thinking about these constructs (Box 1. (2)).

Alternatives to race—for example ethnicity—often seem to come to be used and understood in the same way as race (Condit 2007; Oppenheimer 2001, Sankar and Cho 2002). There is some evidence of a similar definition slippage with respect to 'ancestry' in our dataset, most insidiously where inferred populations labels assigned through genetic ancestry assessment were referred to as races or ethnicities (Table 7). Again, requiring the definition of race, ethnicity and ancestry by authors would highlight their differing utility in addressing different biomedical questions, and assist in prising apart conflation between social and genetic identity.

No article in our sample set discussed ethical or social implications of the reported research (Box 1. (4)), despite recent evidence suggesting geneticists are sensitive to these issues (Ali-Khan and Daar 2010; Caulfield et al. 2009; Lee et al. 2008; Smart et al. 2006). We note that the focus of the six journals from which our study sample was drawn is reporting scientific advances within the field of genetics. Thus, the absence of socio-ethical statements is perhaps not surprising, particularly given that geneticists themselves have emphasized a need for greater awareness and expertise on these issues amongst the authors of genetic studies (Ali-Khan and Daar 2010). Geneticists should consider building their capacity in this area, and/or include such experts on research teams. An important alternative, or in addition to requiring such statements by authors within research articles, would be the regular commissioning by genetics research journals of opinion and review articles by social scientists on the socio-ethical implications of recent genetics advances. In addition, we note a more recent article which provides an example of how socio-ethical concerns can be considered in implementing and reporting genetic studies (Patterson et al. 2010).

About a third of articles did not provide a justification for why they studied the particular research population, or how stratifying by race, ethnicity, ancestry etc., was relevant to the hypothesis under investigation. Notably, articles using race/ethnicity were more likely to specify this information. While this is heartening with respect to the uptake of guidelines for race/ethnicity, it suggests there should be explicit discussion and extension of these to address the uses of 'ancestry'. We note that many of the articles that did not state the reason for the use of a

particular population, ostensibly used the samples for practical reasons not directly related to the research hypothesis—because they were available. In such cases, noting that 'available samples were of 'X' origin' where applicable would contribute to the transparency of the reported research, and may minimize the possibility for misinterpretation vis-à-vis the relationship between genetic and social identity.

Various biomedical journals endorse different combinations of guidelines regarding race/ethnicity, culture, and nationality. However, many do not emphasize them in their online instructions to authors. Journal editors, as gatekeepers of publication standard, seem the intuitive choice to impose such requirements on authors. However, evidence suggests that editors do not feel qualified to develop and apply concrete rules with respect to race and ethnicity (Bhopal et al. 1997; Smart et al. 2006). More importantly, careful consideration of population naming, measurement and definition should occur during study design and research participant recruitment, not ad hoc. A lack of standards on application, definition, classification and measurement of race, ethnicity and ancestry within the genetics community has been noted (Royal et al. 2010; Smart et al. 2006), and personal communication from Dr. Steve Scherer. Such guidelines and standards should be most effective if they are generated through widespread consensus by the genetics community itself. Despite the attention directed to the use and reporting of populations in biomedical study over the last 15 years, our analysis suggests there is still an urgent need for the explicit engagement of these issues by geneticists ((Box 1. (5)), and in particular, to extend the discussion to the uses of ancestry and genetic ancestry ((Box 1. (1), (2), (3), (5)). This might be best achieved through the formation of a dedicated working group including representatives from the spectrum of countries and cultures. Such a group should spearhead discussion of extant guidelines, highlighting their importance for both scientific and socio-ethical reasons, and perhaps their revision and extension in light of the findings of the current study. Broad agreement on, and endorsement of guidelines by the genetics community globally would be a fundamental step forward. Such standards/requirements must then be supported, and consistently enforced by biomedical journal editors (Box 1. (6)).

Acknowledgments This project was funded by Genome Canada through the Ontario Genomics Institute and the Ontario Research Fund—(Genome Canada Competition III), and supported by the McLaughlin-Rotman Centre for Global Health, an academic centre at the University Health Network and University of Toronto. We also sincerely thank Drs. Steve Scherer, Christian Marshall, Jim Lavery, and Ms. Billie-Jo Hardy for helpful conversations through the development of this study. We also thank our anonymous reviewers for their constructive comments.

References

Ali-Khan SE, Daar AS (2010) Admixture mapping: from paradigms of race and ethnicity, to population history. Hugo J

American Academy of Pediatrics: Committee on Pediatric Research (2000) Race/ethnicity, gender, socioeconomic status—research exploring their effects on child health: a subject review. Pediatrics 105:1349–1351

American Anthropological Association (2000) Statement on "Race"

Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 40:616–622

Anonymous (2002) Slicing soup. Nat Biotechnol 20:637

Anonymous (2003) Use of race and ethnicity in public health surveillance: summary of the CDC/ATSDR workshop. Morb Mortal Wkly Rep 42:16

Anonymous (2005) Author guidelines: race, ethnicity and nationality. Paediatric and perinatal epidemiology. In: Wiley-Blackwell. http://www.wiley.com/bw/submit.asp?ref=0269-5022. Accessed 21 Sept 2010

Bamshad MJ, Olson SE (2003) Does race exist? Sci Am 289:78–85

Bamshad M, Wooding S, Salisbury BA, Stephens JC (2004) Deconstructing the relationship between genetics and race. Nat Rev Genet 5:598–609

Bhopal R (1997) Is research into ethnicity and health racist, unsound, or important science? BMJ 314:1751–1756

Bhopal R, Kohli H, Rankin J (1997) Editors' practice and views on terminology in ethnicity and health research. Ethn Health 2:223–227

Bishop DT, Demenais F, Iles MM, Harland M, Taylor JC, Corda E, Randerson-Moor J, Aitken JF, Avril MF, Azizi E, Bakker B, Bianchi-Scarra G, Bressac-de Paillerets B, Calista D, Cannon-Albright LA, Chin-A-Woeng T, Debniak T, Galore-Haskel G, Ghiorzo P, Gut I, Hansson J, Hocevar M, Hoiom V, Hopper JL, Ingvar C, Kanetsky PA, Kefford RF, Landi MT, Lang J, Lubinski J, Mackie R, Malvehy J, Mann GJ, Martin NG, Montgomery GW, van Nieuwpoort FA, Novakovic S, Olsson H, Puig S, Weiss M, van Workum W, Zelenika D, Brown KM, Goldstein AM, Gillanders EM, Boland A, Galan P, Elder DE, Gruis NA, Hayward NK, Lathrop GM, Barrett JH, Bishop JA (2009) Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet 41:920–925

Bonham VL, Warshauer-Baker E, Collins FS (2005) Race and ethnicity in the genome era: the complexity of the constructs. Am Psychol 60:9–15

Braun L, Fausto-Sterling A, Fullwiley D, Hammonds EM, Nelson A, Quivers W, Reverby SM, Shields AE (2007) Racial categories in medical practice: how useful are they? PLoS Med 4:e271

Bronstein M, Pisante A, Yakir B, Darvasi A (2008) Type 2 diabetes susceptibility loci in the Ashkenazi Jewish population. Hum Genet 124:101–104

Brown M (2007) Defining human differences in biomedicine. PLoS Med 4:e288

Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, Mountain JL, Perez-Stable EJ, Sheppard D, Risch N (2003) The

importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 348:1170–1175

Burgner D, Davila S, Breunis WB, Ng SB, Li Y, Bonnard C, Ling L, Wright VJ, Thalamuthu A, Odam M, Shimizu C, Burns JC, Levin M, Kuijpers TW, Hibberd ML, International Kawasaki Disease Genetics Consortium (2009) A genome-wide association study identifies novel and functionally related susceptibility Loci for Kawasaki disease. PLoS Genet 5:e1000319

Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598–604

Caulfield T, Fullerton SM, Ali-Khan SE, Arbour L, Burchard EG, Cooper RS, Hardy BJ, Harry S, Hyde-Lay R, Kahn J, Kittles R, Koenig BA, Lee SS, Malinowski M, Ravitsky V, Sankar P, Scherer SW, Seguin B, Shickle D, Suarez-Kurtz G, Daar AS (2009) Race and ancestry in biomedical research: exploring the challenges. Genome Med 1:8

Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, Froguel P, Balding D, Scott J, Kooner JS (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet 40:716–718

Choudhry S, Taub M, Mei R, Rodriguez-Santana J, Rodriguez-Cintron W, Shriver MD, Ziv E, Risch NJ, Burchard EG (2008) Genome-wide screen for asthma in Puerto Ricans: evidence for association with 5q23 region. Hum Genet 123:455–468

Clayton EW (2002) The complex relationship of genetics, groups, and health: what it means for public health. J Law Med Ethics 30:290–297

Collins FS (2004) What we do and don't know about 'race', 'ethnicity', genetics and health at the dawn of the genome era. Nat Genet 36:S13–S15

Collins FS (2010) The language of life: DNA and the revolution in personalized medicine. Harper-Collins, New York

Comstock RD, Castillo EM, Lindsay SP (2004) Four-year review of the use of race and ethnicity in epidemiologic and public health research. Am J Epidemiol 159:611–619

Condit CM (2007) How geneticists can help reporters to get their story right. Nat Rev Genet 8:815–820

Cooper RS, Kaufman JS, Ward R (2003) Race and genomics. N Engl J Med 348:1166–1170

Crosslin DR, Shah SH, Nelson SC, Haynes CS, Connelly JJ, Gadson S, Goldschmidt-Clermont PJ, Vance JM, Rose J, Granger CB, Seo D, Gregory SG, Kraus WE, Hauser ER (2009) Genetic effects in the leukotriene biosynthesis pathway and association with atherosclerosis. Hum Genet 125:217–229

Davis MM, Bruckman D, Cabana MD, Clark SJ, Dombkowski KJ, Kemper AR, Rushton JL, Freed GL (2001) Constructive use of race and ethnicity variables. Arch Pediatr Adolesc Med 155:973 (author reply 973–974)

Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K, Kim C, Geiger EA, McGrady PW, Blakemore AI, London WB, Shaikh TH, Bradfield J, Grant SF, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM (2009) Copy number variation at 1q21.1 associated with neuroblastoma. Nature 459:987–991

Duster T (2005) Medicine. Race and reification in science. Science 307:1050–1051

Editorial (1996) Ethnicity, race, and culture: guidelines for research, audit, and publication. BMJ 312:1094

Editorial (2004a) The unexamined population. Nat Genet 36:S3

Editorial (2004b) The unexamined 'Caucasian'. Nat Genet 36:541

Foster MW, Sharp RR (2004) Beyond race: towards a whole-genome perspective on human populations and genetic variation. Nat Rev Genet 5:790–796

Ganesh SK, Zakai NA, van Rooij FJ, Soranzo N, Smith AV, Nalls MA, Chen MH, Kottgen A, Glazer NL, Dehghan A, Kuhnel B, Aspelund T, Yang Q, Tanaka T, Jaffe A, Bis JC, Verwoert GC, Teumer A, Fox CS, Guralnik JM, Ehret GB, Rice K, Felix JF, Rendon A, Eiriksdottir G, Levy D, Patel KV, Boerwinkle E, Rotter JI, Hofman A, Sambrook JG, Hernandez DG, Zheng G, Bandinelli S, Singleton AB, Coresh J, Lumley T, Uitterlinden AG, Vangils JM, Launer LJ, Cupples LA, Oostra BA, Zwaginga JJ, Ouwehand WH, Thein SL, Meisinger C, Deloukas P, Nauck M, Spector TD, Gieger C, Gudnason V, van Duijn CM, Psaty BM, Ferrucci L, Chakravarti A, Greinacher A, O'Donnell CJ, Witteman JC, Furth S, Cushman M, Harris TB, Lin JP (2009) Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet 41:1191–1198

Garner CP, Ding YC, John EM, Ingles SA, Olopade OI, Huo D, Adebamowo C, Ogundiran T, Neuhausen SL (2008) Genetic variation in IGFBP2 and IGFBP5 is associated with breast cancer in populations of African descent. Hum Genet 123: 247–255

Geiger HJ (2003) Racial and ethnic disparities in diagnosis and treatment: a review of the evidence and a consideration of causes. In: Smedley B, Stith A, Nelson A (eds) Unequal treatment: confronting racial and ethnic disparities in health care. The National Academic Press, Washington, pp 417–454

Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, Game RM, Rudd DS, Zurawiecki D, McDougle CJ, Davis LK, Miller J, Posey DJ, Michaels S, Kolevzon A, Silverman JM, Bernier R, Levy SE, Schultz RT, Dawson G, Owley T, McMahon WM, Wassink TH, Sweeney JA, Nurnberger JI, Coon H, Sutcliffe JS, Minshew NJ, Grant SF, Bucan M, Cook EH, Buxbaum JD, Devlin B, Schellenberg GD, Hakonarson H (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459:569–573

Gould SJ (1981) The mis measure of man. W. W. Norton & Company, New York

Hancock DB, Romieu I, Shi M, Sienra-Monge JJ, Wu H, Chiu GY, Li H, del Rio-Navarro BE, Willis-Owen SA, Weiss ST, Raby BA, Gao H, Eng C, Chapela R, Burchard EG, Tang H, Sullivan PF, London SJ (2009) Genome-wide association study implicates chromosome 9q21.31 as a susceptibility locus for asthma in Mexican children. PLoS Genet 5:e1000623

Harrison FV (1995) The persistent power of "race" in the cultural and political economy of racism. Annu Rev Anthropol 24:47–74

HUGO Pan-Asian SNP Consortium, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen CH, Chen J, Chen YT, Chu J, Cutiongco-de la Paz EM, De Ungria MC, Delfin FC, Edo J, Fuchareon S, Ghang H, Gojobori T, Han J, Ho SF, Hoh BP, Huang W, Inoko H, Jha P, Jinam TA, Jin L, Jung J, Kangwanpong D, Kampuansai J, Kennedy GC, Khurana P, Kim HL, Kim K, Kim S, Kim WY, Kimm K, Kimura R, Koike T, Kulawonganunchai S, Kumar V, Lai PS, Lee JY, Lee S, Liu ET, Majumder PP, Mandapati KK, Marzuki S, Mitchell W, Mukerji M, Naritomi K, Ngamphiw C, Niikawa N, Nishida N, Oh B, Oh S, Ohashi J, Oka A, Ong R, Padilla CD, Palittapongarnpim P, Perdigon HB, Phipps ME, Png E, Sakaki Y, Salvador JM, Sandraling Y, Scaria V, Seielstad M, Sidek MR, Sinha A, Srikummool M, Sudoyo H, Sugano S, Suryadi H, Suzuki Y, Tabbada KA, Tan A, Tokunaga K, Tongsima S, Villamor LP, Wang E, Wang Y, Wang H, Wu JY, Xiao H, Xu S, Yang JO, Shugart YY, Yoo HS, Yuan W, Zhao G, Zilfalil BA, Indian Genome Variation Consortium (2009) Mapping human genetic diversity in Asia. Science 326:1541–1545

International Council of Medical Journal Editors (2010) Uniform requirements for manuscripts submitted to Biomedical Journals:

writing and editing for biomedical publication, publication ethics: sponsorship, authorship, and accountability. http://www.icmje.org/urm_full.pdf. Accessed 21 Sept 2010

Ioannidis JP, Ntzani EE, Trikalinos TA (2004) 'Racial' differences in genetic effects for complex diseases. Nat Genet 36:1312–1318

Iverson C, Flanagin A, Fontanarosa PB, Glass RM, Giltman P, Lantz JC, Meyer HS, Smith JM, Winker MA, Young RK (1998) American medical association manual of style: a guide for authors and editors, 9th edn

Jenkins G, Merz JF, Sankar P (2005) A qualitative study of women's views on medical confidentiality. J Med Ethics 31:499–504

Kalow W (2001) Pharmacogenetics, pharmacogenomics, and pharmacobiology. Clin Pharmacol Ther 70:1–4

Kaplan JB, Bennett T (2003) Use of race and ethnicity in biomedical publication. JAMA 289:2709–2716

Keyser C, Bouakaze C, Crubezy E, Nikolaev VG, Montagnon D, Reis T, Ludes B (2009) Ancient DNA provides new insights into the history of south Siberian Kurgan people. Hum Genet 126: 395–410

Kressin NR, Chang BH, Hendricks A, Kazis LE (2003) Agreement between administrative data and patients' self-reports of race/ethnicity. Am J Public Health 93:1734–1739

Lee SS (2004) "Incidental findings" of race in pharmacogenomics and the infrastructure for finding differences in biomedical research, pp 81–94

Lee SS (2005) Racializing drug design: implications of pharmacogenomics for health disparities. Am J Public Health 95:2133–2138

Lee SS, Mountain J, Koenig B, Altman R, Brown M, Camarillo A, Cavalli-Sforza L, Cho M, Eberhardt J, Feldman M, Ford R, Greely H, King R, Markus H, Satz D, Snipp M, Steele C, Underhill P (2008) The ethics of characterizing difference: guiding principles on using racial categories in human genetics. Genome Biol 9:404

Lei SF, Yang TL, Tan LJ, Chen XD, Guo Y, Guo YF, Zhang L, Liu XG, Yan H, Pan F, Zhang ZX, Peng YM, Zhou Q, He LN, Zhu XZ, Cheng J, Liu YZ, Papasian CJ, Deng HW (2009) Genome-wide association scan for stature in Chinese: evidence for ethnic specific loci. Hum Genet 125:1–9

Lewontin RC (1995) Human diversity (Scientific American Library Series). W. H. Freeman & Company, second printing edition (October 1995)

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104

Long JC, Kittles RA (2009) Human genetic diversity and the nonexistence of biological races. 2003. Hum Biol 81:777–798

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36:512–517

Nath SK, Han S, Kim-Howard X, Kelly JA, Viswanathan P, Gilkeson GS, Chen W, Zhu C, McEver RP, Kimberly RP, Alarcon-Riquelme ME, Vyse TJ, Li QZ, Wakeland EK, Merrill JT, James JA, Kaufman KM, Guthridge JM, Harley JB (2008) A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. Nat Genet 40:152–154

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456:98–101

Oppenheimer GM (2001) Paradigm lost: race, ethnicity, and the search for a new population taxonomy. Am J Public Health 91:1049–1055

Osborne NG, Feit MD (1992) The use of race in medical research. JAMA 267:275–279

Patterson N, Petersen DC, van der Ross RE, Sudoyo H, Glashoff RH, Marzuki S, Reich D, Hayes VM (2010) Genetic structure of a unique admixed population: implications for medical research. Hum Mol Genet 19:411–419

Plaisier CL, Horvath S, Huertas-Vazquez A, Cruz-Bautista I, Herrera MF, Tusie-Luna T, Aguilar-Salinas C, Pajukanta P (2009) A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. PLoS Genet 5:e1000642

Provine WB (1973) Geneticists and the biology of race crossing. Science 182:790–796

Race, Ethnicity and Genetics Working Group (2005) The use of racial, ethnic, and ancestral categories in human genetics research. Am J Hum Genet 77:519–532

Rasmussen-Torvik LJ, Pankow JS, Jacobs DR Jr, Steinberger J, Moran A, Sinaiko AR (2009) The association of SNPs in ADIPOQ, ADIPOR1, and ADIPOR2 with insulin sensitivity in a cohort of adolescents and their parents. Hum Genet 125:21–28

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature 461:489–494

Risch N, Burchard E, Ziv E, Tang H (2002) Categorization of humans in biomedical research: genes, race and disease. Genome Biol 3 (comment 2007)

Rivara F, Finberg L (2001) Use of the terms race and ethnicity. Arch Pediatr Adolesc Med 155:119

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Rotimi CN (2004) Are medical and nonmedical uses of large-scale genomic markers conflating genetics and 'race'? Nat Genet 36:S43–S47

Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, Clark AG (2010) Inferring genetic ancestry: opportunities, challenges, and implications. Am J Hum Genet 86:661–673

Rung J, Cauchi S, Albrechtsen A, Shen L, Rocheleau G, Cavalcanti-Proenca C, Bacot F, Balkau B, Belisle A, Borch-Johnsen K, Charpentier G, Dina C, Durand E, Elliott P, Hadjadj S, Jarvelin MR, Laitinen J, Lauritzen T, Marre M, Mazur A, Meyre D, Montpetit A, Pisinger C, Posner B, Poulsen P, Pouta A, Prentki M, Ribel-Madsen R, Ruokonen A, Sandbaek A, Serre D, Tichet J, Vaxillaire M, Wojtaszewski JF, Vaag A, Hansen T, Polychronakos C, Pedersen O, Froguel P, Sladek R (2009) Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. Nat Genet 41:1110–1115

Sankar P, Cho MK (2002) Genetics. Toward a new vocabulary of human genetic variation. Science 298:1337–1338

Sankar P, Cho MK, Condit CM, Hunt LM, Koenig B, Marshall P, Lee SS, Spicer P (2004) Genetic research and health disparities. JAMA 291:2985–2989

Sankar P, Cho MK, Mountain J (2007) Race and ethnicity in genetic research. Am J Med Genet A 143A:961–970

Schwartz RS (2001) Racial profiling in medical research. N Engl J Med 344:1392–1393

Shanawani H, Dame L, Schwartz DA, Cook-Deegan R (2006) Non-reporting and inconsistent reporting of race and ethnicity in articles that claim associations among genotype, outcome, and race or ethnicity. J Med Ethics 32:724–728

Shi H, Tan SJ, Zhong H, Hu W, Levine A, Xiao CJ, Peng Y, Qi XB, Shou WH, Ma RL, Li Y, Su B, Lu X (2009) Winter temperature and UV are tightly linked to genetic changes in the p53 tumor suppressor pathway in Eastern Asia. Am J Hum Genet 84:534–541

Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW (2004) The genomic distribution of

population substructure in four populations using 8, 525 autosomal SNPs. Hum Genomics 1:274–286

Smart A, Tutton R, Ashcroft R, Martin PA, Ellison GTH (2006) Can science alone improve the measurement and communication of race and ethnicity in genetic research? Exploring the strategies proposed by nature genetics. BioSocieties 1:313–324

Stevens J (2003) Racial meanings and scientific methods: changing policies for NIH-sponsored publications reporting human variation. J Health Polit Policy Law 28:1033–1088

Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, Schork NJ, Risch NJ (2005) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 76:268–275

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044

Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, Willman C, Neale G, Downing J, Raimondi SC, Pui CH, Evans WE, Relling MV (2009) Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet 41:1001–1005

Via M, Ziv E, Burchard EG (2009) Recent advances of genetic ancestry testing in biomedical research and direct to consumer testing. Clin Genet 76:225–235

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320:539–543

Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, Salyakina D, Imielinski M, Bradfield JP, Sleiman PM, Kim CE, Hou C, Frackelton E, Chiavacci R, Takahashi N, Sakurai T, Rappaport E, Lajonchere CM, Munson J, Estes A, Korvatska O, Piven J, Sonnenblick LI, Alvarez Retuerto AI, Herman EI, Dong H, Hutman T, Sigman M, Ozonoff S, Klin A, Owley T, Sweeney JA, Brune CW, Cantor RM, Bernier R, Gilbert JR, Cuccaro ML, McMahon WM, Miller J, State MW, Wassink TH, Coon H, Levy SE, Schultz RT, Nurnberger JI, Haines JL, Sutcliffe JS, Cook EH, Minshew NJ, Buxbaum JD, Dawson G, Grant SF, Geschwind DH, Pericak-Vance MA, Schellenberg GD, Hakonarson H (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459:528–533

Wing MR, Ziegler J, Langefeld CD, Ng MC, Haffner SM, Norris JM, Goodarzi MO, Bowden DW (2009) Analysis of FTO gene variants with measures of obesity and glucose homeostasis in the IRAS family study. Hum Genet 125:615–626

Winker MA (2004) Measuring race and ethnicity: why and how? JAMA 292:1612–1614

Wood AJ (2001) Racial differences in the response to drugs—pointers to genetic differences. N Engl J Med 344:1394–1396

Xiong DH, Liu XG, Guo YF, Tan LJ, Wang L, Sha BY, Tang ZH, Pan F, Yang TL, Chen XD, Lei SF, Yerges LM, Zhu XZ, Wheeler VW, Patrick AL, Bunker CH, Guo Y, Yan H, Pei YF, Zhang YP, Levy S, Papasian CJ, Xiao P, Lundberg YW, Recker RR, Liu YZ, Liu YJ, Zmuda JM, Deng HW (2009) Genome-wide association and follow-up replication studies identified ADAMTS18 and TGFBR3 as bone mass candidate genes in different ethnic groups. Am J Hum Genet 84:388–398

Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, Nakamura Y, Kamatani N (2008) Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. Am J Hum Genet 83:445–456

Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y, Yamagata K, Hinokio Y, Wang HY, Tanahashi T, Nakamura N, Oka Y, Iwasaki N, Iwamoto Y, Yamada Y, Seino Y, Maegawa H, Kashiwagi A, Takeda J, Maeda E, Shin HD, Cho YM, Park KS, Lee HK, Ng MC, Ma RC, So WY, Chan JC, Lyssenko V, Tuomi T, Nilsson P, Groop L, Kamatani N, Sekine A, Nakamura Y, Yamamoto K, Yoshida T, Tokunaga K, Itakura M, Makino H, Nanjo K, Kadowaki T, Kasuga M (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 40:1092–1097

Yeager M, Chatterjee N, Ciampa J, Jacobs KB, Gonzalez-Bosquet J, Hayes RB, Kraft P, Wacholder S, Orr N, Berndt S, Yu K, Hutchinson A, Wang Z, Amundadottir L, Feigelson HS, Thun MJ, Diver WR, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Crawford ED, Haiman CA, Henderson B, Kolonel L, Le Marchand L, Siddiq A, Riboli E, Key TJ, Kaaks R, Isaacs W, Isaacs S, Wiley KE, Gronberg H, Wiklund F, Stattin P, Xu J, Zheng SL, Sun J, Vatten LJ, Hveem K, Kumle M, Tucker M, Gerhard DS, Hoover RN, Fraumeni JF Jr, Hunter DJ, Thomas G, Chanock SJ (2009) Identification of a new prostate cancer susceptibility locus on chromosome 8q24. Nat Genet 41: 1055–1057