Different differences: The use of 'genetic ancestry' versus race in biomedical human genetic research

Joan H. Fujimura and Ramya Rajagopalan

Social Studies of Science 2011 41: 5 originally published online 7 December 2010
DOI: 10.1177/0306312710379170

S|S|S

Social Studies of Science
41(1) 5–30

DOI: 10.1177/0306312710379170
sss.sagepub.com
SAGE

Different differences: The use of 'genetic ancestry' versus race in biomedical human genetic research

Joan H. Fujimura
University of Wisconsin, USA

Ramya Rajagopalan
University of Wisconsin, USA

Abstract

This article presents findings from our ethnographic research on biomedical scientists' studies of human genetic variation and common complex disease. We examine the socio-material work involved in genome-wide association studies (GWAS) and discuss whether, how, and when notions of race and ethnicity are or are not used. We analyze how researchers produce simultaneously different kinds of populations and population differences. Although many geneticists use race in their analyses, we find some who have invented a statistical genetics method and associated software that they use specifically to avoid using categories of race in their genetic analysis. Their method allows them to operationalize their concept of 'genetic ancestry' without resorting to notions of race and ethnicity. We focus on the construction and implementation of the software's algorithms, and discuss the consequences and implications of the software technology for debates and policies around the use of race in genetics research. We also demonstrate that the production and use of their method involves a dynamic and fluid assemblage of actors in various disciplines responding to disciplinary and sociopolitical contexts and concerns. This assemblage also includes particular discourses on human history and geography as they become entangled with research on genetic markers and disease. We introduce the concept of 'genome geography' to analyze how some researchers studying human genetic variation 'locate' stretches of DNA in different places and times. The concept of genetic ancestry and the practice of genome geography rely on old discourses, but they also incorporate new technologies, infrastructures, and political and scientific commitments. Some of these new technologies provide opportunities to change some of our institutional and cultural forms and frames around notions of difference and similarity. Nevertheless, we also highlight the slipperiness of genome geography and the tenacity of race and race concepts.

Keywords

ancestry, biomedical genetics, ethnography, genome geography, population, race, socio-material practice

Corresponding author:
Joan H. Fujimura, Department of Sociology, University of Wisconsin-Madison, 8128 Social Science Building, 1180 Observatory Drive, Madison, WI 53706, USA.
Email: fujimura@ssc.wisc.edu

This article analyzes the role played by variable concepts of population and population differences in biomedical genomic studies, and examines how geneticists construct and use different conceptions of population. In particular, it examines the role of the population designators 'race' and 'genetic ancestry' and the relationship between the two concepts in 21st century biomedical genetic research. In their search for biological contributions to complex diseases, geneticists have been identifying parts of genomic DNA that they believe are medically relevant and that are passed down and inherited across generations. But the inheritance of DNA implies ancestors and descendents, which in contemporary biomedical genomics have been operationalized as groups of people thought to share similar genetics. Despite the widely held view that races are complex sociocultural entities rather than genetically bounded categories, some studies of genetics and disease (for example, Burchard et al., 2003) explicitly call on concepts of race in their analysis of genetic data. Such studies have generated criticism and debate amongst geneticists, bioethicists, social scientists, and public health scholars during the last 15 years.[1] They have noted that there might seem to be some propinquity between biology and race, because of the ways racial discrimination, social stratification, health disparities, and mistrust of the scientific establishment feed back into the health of individuals (for example, Duster, 2003; Montoya, 2007; Rebbeck and Sankar, 2005; Shields et al., 2005). This overlap of socio-historical and genetically bounded categories, rather than ushering in an era of personalized medicine tailored to *individuals* with perceived differential genetic risks, could lead instead to a situation where medical diagnosis and care are organized around already socially stratified *groups* with perceived differential genetic and health risks.

Our larger ethnographic project explores how population is conceived and operationalized in five human genetic variation and disease projects. This article speaks specifically to issues that have been subjects of debate in the world of race, genetics, disease, and health policy. Our study has found that *some* biomedical geneticists who conduct genome-wide association studies (GWAS) agree with critiques of the use of race in the search for genetic contributions to complex diseases. In particular, these geneticists agree with the argument that race categories do not correspond to genetic groupings and therefore are not appropriate tools to search for disease-related genes. Although these GWAS researchers forego the use of race, they argue that their large-scale genome-wide scans for disease associations require methods for accounting for *population* differences. These researchers have attempted to develop tools that they believe to be more accurate than race categories for accounting for population differences. This article focuses on the socio-material construction of these tools.[2]

The medical geneticists we studied have developed new methods for measuring genetic similarities and differences using genomic and statistical technologies. Although they refer to the samples that their software groups together as 'genetically similar', some downstream and upstream users of the technology have used the term 'ancestry' to refer to the population differences that these technologies were designed to measure. For example, some of the researchers we studied invoked concepts of 'genetic ancestry' when discussing the usefulness of these new technologies to their search for gene–disease associations.

The technologies used in making these assessments of 'genetic ancestry' differ from the admixture mapping and 'ancestry informative markers' (AIMs) technologies used in

commercial genetic genealogy tests, which usually assess ancestry based on continental race categories. Still, we find that some audiences read genetic ancestry as synonymous with race. Such readings stem in part from the practices of what we call 'genome geography', which operate in the research designs and the technologies used to construct populations. 'Genome geography' refers to how, through the tools and practices of human genetics, bits of genomic sequence become associated with specific geographic locations, posited as the place of origin of people who possess these bits. For example, some studies use GWAS technologies to postulate the geographic origins of contemporary peoples, and this inference of 'shared ancestry' is understood by some audiences as isomorphic with race and ethnic categories.

Genome geography then is one thread between population, race, and 'genetic ancestry' that renders the three concepts difficult to untangle. Although one can interpret genetic ancestry as a concept derived from population genetics technologies and race as a sociocultural set of understandings, we show that the two are not so easily separated in scientific/cultural practices and discourses. We analyze how these terms are used (and in some instances measured) in new genomics studies of disease.

We first briefly discuss our methods, before presenting ethnographic data on the construction and implementation of EIGENSTRAT, a population genetics software technology. We illustrate how some biomedical researchers use EIGENSTRAT to avoid emphasizing populations in their search for disease-related DNA and certainly to avoid the use of race. We also show, however, how other researchers using EIGENSTRAT find it difficult to give up on geographically 'locating' DNA and designating populations; that is, they move from *genetic similarity* to *genetic ancestry* to *genome geography*. We end with a discussion of potential consequences of EIGENSTRAT and of genome geography.

Methods

The present article is based on a larger ethnographic study of the use of the seemingly self-evident notion of *population* in the design, data collection, and analysis stages of human genetic variation studies. In order to study the practices through which populations and population differences are constructed and used in biomedical research, we conducted ethnographic research in laboratories and at scientific meetings, and documentary analyses of published literature on genetic variation and of federal regulations that apply to human genetic variation research and that guide scientists in choosing or describing subjects.

This article focuses on the production and consumption of a software package and its incorporated algorithms as used in human genetic variation and biomedical genetics studies. Ethnographic analyses of the *construction* of software technologies and their implications include, for example, Suchman's (2007) studies of artificial intelligence, but have not been done in the field of genetics.[3]

This article is based on interviews and field research in and around three major medical genetics research centers, each of which conducts research with other laboratories and researchers. The field research was conducted over a period of 2.5 years, starting in January 2007. We do not use names of quoted research subjects as part of our agreements

to protect our human subjects. We also note that this article does not aim to represent all views of all geneticists interviewed in our larger study, but aims to make particular points based on the data.

GWAS

Recently, new genomic technologies[4] have enabled researchers to conduct GWAS to search for genetic markers that may increase the risk of developing common complex diseases (CCD). GWAS methods have become widely and quickly adopted, as evidenced by a sharp spike in publications on the subject over the past 2 years. In December 2007, the journal *Science* named human GWAS its 'breakthrough of the year' (Pennisi, 2007). Hundreds of genetic markers have since been reported by researchers as being associated with more than 80 diseases and traits (Hindorff et al., 2009).

Although GWAS researchers state that their methods can be used to study the genetics of any trait, they typically seek genetic markers that may increase the risk of heart disease, type 2 diabetes, cancer, and other CCD. These diseases are thought to be common because they are not restricted to particular families or populations.

Their etiology is complex because they manifest for different reasons in different individuals. These are believed to include heterogeneous combinations of 'environmental' factors such as diet, exercise, and toxic exposure, coupled with contributions from potential genetic risk factors in an individual's genome. For these reasons, complex diseases are more difficult to study than simpler genetic diseases, such as those caused primarily by a single gene whereby a person who has a disease-associated version of the implicated gene will almost certainly show symptoms of the disease. GWAS researchers focus on the genetic elements of CCD causation, although they acknowledge that many other factors and processes are involved in the etiology of CCD.[5] More recent GWAS, including some at our study sites, are beginning to examine environmental factors in disease through investigations of gene-environment interactions.[6]

GWAS researchers scan the genomes of large groups of individuals in search of genetic markers associated with CCD. They identify such markers through statistical analyses that distinguish differences in the frequencies of genetic marker variants in cases afflicted with a disease versus controls. The strategy is to genotype[7] and compare hundreds of thousands of genetic markers between cases and controls, to find differences statistically associated with the cases.

This statistical approach is an exploratory strategy that researchers consider to be 'hypothesis-generating', rather than hypothesis-driven. An example of the latter is candidate gene studies, where scientists look at a small set of genes they suspect to be involved in a disease. According to our respondents, candidate gene studies examine only a few genes at a time and rely on prior information that indicates that those genes might be related to a disease. They argue that, in contrast, GWAS survey the genome, in both 'known genes' and in the vast regions of the genome that lie outside of known genes. The markers themselves often lie outside of known genes, or in regions whose function remains 'uncharacterized', which initially led other researchers to question the relevance of these studies. Indeed, these large and expensive GWAS generated some controversy when they were under development. In response, GWAS researchers argued that in the absence of a priori information, genome-wide searches were

necessary to identify regions of the genome involved in complex diseases. They argued that locating disease-associated genetic markers, including those outside known genes, would be a first step towards pinpointing regions of the genome involved in complex diseases, which could then be examined in greater detail using 'wet lab' molecular and cell biology techniques.

GWAS are large, expensive, and time-consuming because they attempt to study genetics across entire genomes. The success of the statistical analysis depends on a high degree of 'power', that is, a high chance of detecting a marker associated with disease. The researchers increase power by increasing both the number of individuals sampled and the number of markers that are genotyped across each individual's genome. They collect DNA samples from thousands of people, and genotype and evaluate hundreds of thousands of genetic markers in each sample. When they have collected enough[8] cases and controls for a particular disease condition, and genotyped enough markers, researchers statistically analyze potential relationships of genotyped markers to health outcomes.

Genetic variation and concepts of population in GWAS

Once DNA samples are collected and genotyped, GWAS researchers analyze them with a focus on two ways that genetic variation can distinguish their case samples from their control samples. They are interested in identifying genetic marker variants that tend to occur more frequently in the genomes of sampled individuals with the disease than those without, on the assumption that such variants may be related to differential risks of the disease between cases and controls. But some variants that may have nothing to do with the disease under study can also differ between cases and controls. Researchers refer to such variation as indicative of 'population substructure', or different subpopulations within the samples. Substructure may confound the GWAS search for markers related to disease by generating spurious findings, and therefore has been the subject of geneticists' efforts to find ways to identify, account and control for it.

GWAS currently focus on genetic markers called single nucleotide polymorphisms (SNPs).[9] A SNP is a single base pair position in genomic DNA at which two or more possible sequence alternatives (called 'variants') can occur. Researchers genotype each DNA sample, at each marker, to determine the pattern of variants in each case and control individual. Geneticists estimate that there are about 10 million SNPs across the approximately 3 billion bases of the human genome, some of which lie within the approximately 20,000 to 25,000 genes.[10]

GWAS researchers use the *frequencies* of SNP variants in cases versus controls to make their statistical determinations of which SNPs are related to disease. Variant frequencies may only differ between cases and controls at some SNPs. More generally in human genetic variation studies, researchers believe the frequencies of many SNP variants differ across human *populations* – a SNP variant that is common (or high in frequency) in one population may be rare (or at low frequency) in another (Brookes, 1999). For example, a SNP may have two variants, 'A' or 'C'. A given individual will have two copies of the SNP (because each chromosome is present in two copies), and each copy will be either 'A' or 'C'. The frequency is usually expressed as a percentage that says how many times 'A' occurs, relative to 'C' (or vice-versa), in a sample of individuals deemed to be members of the population. Thus, in some groupings of individuals, 'A'

might be more common than 'C', while in other groups, 'C' might be more common. The frequencies will depend on how the groups are constructed.

Part of the job of GWAS is to try to identify frequency differences in SNPs between case and control groups that may indicate increased risk for the particular disease under study, rather than being due to population substructure within the groups. Thus, *notions of genetic variation between populations* are key to GWAS searches for disease-related genetic markers. However, *concepts of population are also key to ideas of genetic variation.* Methods used to measure SNP frequencies are based on concepts of 'population' as well as methods for sampling populations. *So what is 'population'?*

Population is defined in many different ways, depending on the particular study, the discipline, the time period, and so on. In human genetic variation studies, of which GWAS is only one kind, there has been virtually no standardization of population categories; instead, the groups being analyzed are often referred to as 'populations' without clear definition. Many genetic studies delineate populations according to other human group descriptors, such as sample donors' self-identified race or ethnicity, nation of origin or regional identification, continent of ancestry, and so on. Even within GWAS, there is extensive variation around how researchers construct and map genetic variation and populations. Using ethnographic methods, we explored how some GWAS have operationalized and accounted for population and population differences.

Methods for accounting for population substructure in GWAS: EIGENSTRAT

In order to find statistically significant SNPs that appear to increase disease risk, the GWAS researchers we studied developed methods to 'adjust' the 'raw' data to enhance the likelihood that any SNPs identified in their subsequent analysis would be related to the disease and not to 'population substructure'. To minimize the error rate of false positive associations between a SNP and the disease being studied, researchers wanted to account for genetic differences between cases and controls that had nothing to do with the CCD they are studying. They called this 'adjusting for population substructure/stratification'. Population substructure or stratification refers to groups of individuals in a sample who may be more genetically similar to each other than they are to the rest of the sample. GWAS researchers believe that accounting for population substructure reduces 'noise' in the data and makes their control samples more 'comparable' to their case samples.[11] Many of these SNP variants shared by the case samples may not be disease-related, so researchers take steps to prevent these SNPs from being incorrectly identified as associated with the disease.

Instead of designing their genetic analysis by relying on a priori designations of population categories, such as pre-existing race or nationality labels, the medical geneticists we studied used a statistical procedure that they regard as more accurate for adjusting for population substructure. That is, although some of their samples may have been collected and labeled using racial or ethnic category designations, these medical geneticists specifically chose not to use such information in their analyses.

Instead, they chose to develop ways to control for substructure by adjusting their genotyping data before conducting statistical analyses of SNP associations to disease. Some of them stated that they believe that racial or ethnic categories do not refer to a genetic set of categories, and therefore, are inappropriate for controlling for substructure

and analyzing genetic associations. Instead, they argued for what a respondent called a 'genetic standpoint':

> It becomes very important to us to know whether those groups actually represent groups of individuals that can be treated genetically as comparable people. So, if we're doing cases and controls, we want to know our cases are comparable to our controls from a genetic standpoint … Because ... if you don't do that properly … you can end up with some very apparently interesting results that actually have nothing to do with biology or genetics, but they just have to do with population differences.

By population differences, this geneticist means differences due to population substructure. The primary tool these medical geneticists used to account for population substructure is a statistical algorithm, EIGENSTRAT, designed by a population genetics research group, including a mathematician, a statistical geneticist and a population geneticist.[12] This algorithm is part of a larger software package, known as Eigensoft, which includes other programs that either generate input for EIGENSTRAT or use its output for other kinds of analyses. The primary input for EIGENSTRAT comes from another program, smartpca, that generates 'axes of variation' throsugh a classical statistical approach called principal components analysis (PCA). The medical geneticists then use these axes in a statistical regression analysis to adjust for population substructure. That is, using EIGENSTRAT, they generate adjusted numerical values along the axes of variation, which summarize an individual's variation across all the SNPs. We will refer to these values as *SNP variation scores*. They then analyze these adjusted values for associations to disease.[13] EIGENSTRAT has been adopted by many other research groups doing GWAS analysis.

The output of EIGENSTRAT is sometimes used to divide a group of samples into similarity groups, or clusters, on the basis of the SNP variation scores. Other programs in the Eigensoft software package were designed to perform this clustering of individuals with similar patterns of SNP variants. These other Eigensoft programs build clusters via algorithms that do not depend on pre-labeling or pre-sorting samples using a priori ideas of human evolution or historical migrations. Nor do they require pre-specifying the number of expected clusters or groups, as does the program *Structure* (cf. Bolnick, 2008) Researchers cannot predetermine the number of clusters that are drawn in the Eigensoft plots, but they can use algorithms in the Eigensoft package to plot each individual's scores for the top axes of variation, and examine resulting plots for clusters of individuals (where a cluster is interpreted as similar marker patterns). Because the Eigensoft programs produce these clusters without any reference to self-reported race, ethnicity or ancestry, they can be used to create *categories of genetic difference* that are *not* categories of race.[14]

Still, when using Eigensoft in studies of disease, researchers do not need to run the EIGENSTRAT output through the clustering programs, nor do they need to produce clusters of any kind, as noted by the methods article that initially described the EIGENSTRAT approach:

> The PCA method does not attempt to classify all individuals into discrete populations or linear combinations of populations, which may not always be the correct model for population history. Instead, PCA outputs each individual's coordinates along axes of variation. An algorithm could

> in principle be used as a post-processing step to cluster individuals based on their coordinates along these axes, but we have not implemented this. (Patterson et al., 2006)

Different uses of EIGENSTRAT

Nevertheless, as a demonstration of the possibility of generating clusters, the authors did include two such plots later in their methods paper. They also simultaneously submitted another paper that included another plot. In that paper, they hypothesized that such plots can be used to represent genetic ancestry differences between, for example, 'northwest Europe' and 'southeast European' subpopulations. Other researchers in later publications have used an Eigensoft to produce cluster plots of individuals within Finland and Sweden, as well as Japan.

The articles with plots that separate subpopulations have primarily been published by population geneticists. This is unsurprising, because demarcating populations and population substructure is central to their field. In contrast, GWAS analysis in medical genetics does not require researchers to plot and separate clusters of individuals based on their genetic variation scores. The medical geneticists we interviewed[15] used the plots as a check of their methods and data, but they have tended not to publish the plots. Some said that there is no need to mention specific ancestries or geographic origins when using EIGENSTRAT in GWAS because EIGENSTRAT automatically adjusts for population substructure using SNP variation scores before disease associations are measured.

Nevertheless, these medical researchers still made plots to check whether graphed EIGENSTRAT SNP variation scores grouped individuals as *expected*. That is, they checked whether these plots group samples according to the researchers' information and assumptions about sampled individuals' ethnic or geographic backgrounds. When showing a plot based on SNP variation scores of samples collected in Finland, a researcher pointed to an 'X' that marked an outlier, someone whose scores put them at a distance outside the cluster of other 'X's', and said that this person's grandmother may have come from another country. In another laboratory, a researcher showed a plot of SNP variation scores of individuals who self-identified as native Hawaiians. Those graphed scores were relatively spread out in an ellipse between HapMap plots of 'Japanese in Tokyo, Japan', 'Tuscans in Italy', 'Yoruba in Ibadan, Nigeria' and 'Maasai in Kinyawa, Kenya'. Both researchers were checking to see whether their program was working correctly by comparing plotted clusters with their expected clusterings. This was an intermediate quality control step. In the first example, the researchers explicitly stated that they had no intention of publishing the plots.

Race versus 'genetic ancestry'

Some of the medical geneticists we interviewed were enthusiastic about the EIGENSTRAT program as the data-organizing tool they chose because it also allowed them to avoid the use of socio-cultural categories such as race. In their view, Eigenstrat allowed them to account for genetic differences in their analyses, while race categories – which they viewed as socio-cultural concepts – were inappropriate for their genetic analyses:

> For my personal conception, race is a social term. It's applied by people in a society, and that's just how they distinguish people …. But ... that's not how I phenotype [my patients] or the participants

> in studies. So, for me as a geneticist ... the participants that I study typically are defined either by their self-described race – white, black, non-Hispanic white, kind of the OMB [Office of Management and Budget] terms that our grants require that we write down, which is a start. But it obviously doesn't capture what I need to know as a geneticist, which is their– the genetic variation that they have. Most of the time, it doesn't matter. Most of the time, whether you have African or Asian or European ancestry. Since we're all African originally, in our ancestry if you go far enough back, then most of the variation is common to all, ah, racial populations or continental population terms ….
>
> I'm circumspect about the role that those social terms have in the research that I do, and I still try to reclassify using genetic information. ... [I'm] agnostic to whether that means that there's self-described more black ancestry or self-described more white ancestry. Again, I don't know what those terms mean, black and white, but if I can distinguish statistically that one group is more similar than another group, genetically, then, it's probably appropriate for me to stratify my analyses and account for those apparent genetic differences.

Some researchers said their tools use 'ancestry' in place of race categories to account for population substructure:

> But, if it's not recognized that there are differences in the, say, ancestry since the last 50,000 years of human history, it can lead to problems. So you could mistake an association between disease and genetic variance that could distinguish, say, continental African versus continental European ancestry or Asian. So, I formally divide my study subjects into self-described categories of people, based on whether they describe themselves as black or white. But I then do tests – genetic tests – that ignore [those self-described categories] and try to cluster what the ancestral backgrounds are of the different individuals .... [I try to do] formal tests to see if I can distinguish different groups of ancestry within a population.

In place of race, some medical geneticists insisted that they use a technical definition of populations, based on genetic similarity. Some of them used terms such as *genetic history* and *ancestry* in addition to or sometimes instead of *genetic similarity*. Some said that they believe that the EIGENSTRAT program allows them to create clusters that 'let the data tell them' the 'genetic history'[16] of individuals who are clustered together:

> You basically just *let the data tell you* – … if you type hundreds of thousands of markers, essentially what you can say is there seem to be three different directions ... in which the *genetic history* of the group of people that you're studying differs. They form kind of a cluster, like this, along one axis, then they form sort of two clusters along another axis, and two clusters along another axis, and that can let you basically assign people into *genetically-similar groups*. [italics added]

We focus here on how scientists assess or construct ancestry using quantitative tools to measure 'genetic ancestry' during the genotyping process. Many of the researchers called the process of using EIGENSTRAT 'adjusting for ancestry', and they viewed this practice as a requirement for assuring that the results they obtain through genetic association studies are revealing disease-marker associations and not measuring population differences.

'Adjusting for ancestry' is another phrase they used to mean 'adjusting for population substructure' or 'adjusting for population stratification':

> First, we apply principal components analysis to genotype data to infer continuous axes of genetic variation. Intuitively, the axes of variation reduce the data to a small number of dimensions, describing as much variability as possible; they are defined as the top eigenvectors of a covariance matrix between samples …. Second, we continuously adjust genotypes and phenotypes by amounts attributable to ancestry along each axis, via computing residuals of linear regressions; intuitively, this creates a virtual set of matched cases and controls. Third, we compute association statistics using ancestry-adjusted genotypes and phenotypes. (Price et al., 2006)

How does this translation or slippage from 'genetically similar groups' to groups that share 'genetic history' and 'ancestry' – at least by some of the GWAS researchers we interviewed – happen? This is the question we address next. We begin here by examining the notion of genetic ancestry. Just as using Eigensoft to adjust for population substructure in GWAS does not require that individuals with similar genetic scores be clustered and plotted, neither does it require any reference to 'ancestry' to do the work of adjusting for population substructure, or 'to genetically match cases and controls'.[17] In EIGENSTRAT the PCA scores and the adjusted SNP variation scores are not pre-sorted into subpopulations at any step, and they do not have to be labeled with any ancestry labels (Price et al., 2006).

What does ancestry mean in the above quote? Like 'population', 'ancestry' is a loose term and has different representations, interpretations, and meanings in different contexts. Ancestry, in some cases, refers to lines of descent, or hierarchical inheritance relationships, as in genealogy trees. In disease association studies such as GWAS, ancestry (a shorthand for genetic ancestry) is conceptualized as a confounding factor that needs to be accounted for in order to detect statistically significant disease-related SNPs.

While the American Society for Human Genetics (ASHG), regarded by geneticists as the leading human genetics association in the USA, notes that ancestry determination depends on 'how underlying patterns of human genetic diversity are distributed among populations' (American Society for Human Genetics, 2008), we argue that *these underlying patterns are known only through the data and data producing technologies of the geneticists*. For example, there are methodological differences in the assessments of a person's ancestry based on how far back in time the assessment goes. According to the ASHG (2008), 'ancestry estimation' differs depending on the time period in which geneticists want to assess ancestry. Very recent generations of ancestors are often assessed through family history information. But the levels of human ancestry that extend further back, while important to researchers of human genetic variation, are also the most controversial in public debates over ancestry determinations and the most difficult to define and interpret with precision.

The EIGENSTRAT designers claim that they do not intend to confer a label of ancestry on the *individual*, since an individual can have multiple ancestries through different combinations of SNPs.[18] Instead, the Eigensoft clustering program is used to assign ancestry only to the combination of SNPs summarized in the coordinates of the top principal components. These coordinates are believed to summarize each individual's SNP patterns. Researchers believe comparing these values tells them how similar an individual

or group of individuals is to others in the sample. Some GWAS researchers then translate this measure of similarity into a particular ancestry depending on where the similarly clustered SNP variation scores fall in relation to other clusters of scores on a plot.

This points to two sets of infrastructures that are used by geneticists to designate genetic ancestry. First, ancestry determination depends critically on which of the approximately 10 million SNPs are selected for genotyping. Scientists who designed the SNP genotyping chips selected which SNPs to include. They did so by relying heavily on data in the International Haplotype Map (HapMap) to select a set of SNPs located throughout the genome that would provide 'reasonable' coverage of the genetic variation in the sampled HapMap populations. In determining the number and types of SNPs for the chips, they took into account population genetic considerations, such as which groups the chips will be most suitable for when assessing genetic variation. They also took into account 'technical' considerations such as which SNPs will 'perform' the best in the genotyping experiments. The building of the chip, the selection of markers, and the genotyping itself are central to GWAS.[19]

Second, genetic ancestry determinations depend on the set of reference samples to which individuals' SNP variation scores are compared in order to label the cluster they fall into. That is, how these reference samples are themselves labeled has implications for how test samples are labeled. Reference samples differ depending on the particular study, but typically rely on a choice as to which population the researchers think is 'genetically' closest to the test samples being analyzed, which usually involves geographic considerations. One paper that used Eigensoft compared samples collected by collaborators in Thailand to reference samples of Chinese and Japanese populations from the HapMap (Patterson et al., 2006). Another article used self-reported nationality as a reference (Price et al., 2008) and concluded that SNP variation scores of genotyped European samples corresponded to a roughly northwest–southeast gradient of variation in Europe that could distinguish different European ancestries. Another study examined African-American men and compared their SNP variation scores with those of Yoruba and Kenyans at one end and Italian-Americans at the another end, to visually depict the variation in genetic ancestry among African-Americans.

To summarize, genetic ancestry is calculated by inference, which is ultimately a statistical exercise that depends on the analytic software itself and the information that goes into the statistical software, including the genetic markers used in genotyping and the reference population data used for comparative purposes.

But is using notions of ancestry in medical genetics problematic? If ancestry is shorthand for designating groups of individuals with similar SNP variation patterns, then the ASHG believes it to be a useful concept. As the ASHG Statement on Ancestry Testing notes, 'For epidemiological purposes, inference of cohesiveness of ancestral history is more relevant than is the specification of particular populations of ancestral origin' (American Society for Human Genetics, 2008). Here, ASHG argues that for studies of disease, researchers could use genetic data to *hypothesize* that a group of people with similar genetics *might* share similar ancestry or genetic history, but need not *make claims* that these individuals are members of a specific labeled population or share a specific labeled ancestry. In other words, for epidemiological purposes, *similar genetics* matter and *ancestry labels* do not matter.

The distinction between similar genetics and ancestry labels matters, as we shall see below. Ancestry labels, for example continental labels such as 'European' or 'African',

tend to group together individuals with similar SNP variants, and then go a step further and attribute the similarities to shared ancestry and descent from peoples from a single geographic locale. On the other hand, ASHG states that similar genetics alone is what matters and is useful. No explanation for why those similarities are present is necessary.

Many of the GWAS researchers we studied agreed with the ASHG statement, and felt that they could operationalize 'similar genetics' in practice. But even some of them found it difficult to maintain the distinction. Slippage occurred when some researchers inferred *shared* ancestry on the basis of EIGENSTRAT SNP variation scores, and then went on to label these ancestries with names that connote ancestral *geographic* origins or *socio-cultural* labels. We discuss this inference process in the next section.

Genome geography: Population genetics and the translation of SNP similarity to relatedness

We have traced the move from 'SNP similarities' to 'genetic history' and 'ancestry'. Except for the first, these concepts imply relatedness and belonging. The population geneticists who designed the EIGENSTRAT tool believe that individuals with similar patterns of genetic markers have related 'genetic histories'. The medical geneticists are comfortable using terms such as 'genetic history' and 'ancestry' when speaking of groups of individuals with similar SNP variation scores, in part because they have adopted the language and ideas used by the population geneticists who designed the technologies they use to correct for population substructure. They have also learned similar ideas through their coursework on human evolution and human genetics.

The move from similarity to relatedness is rooted in theories from population genetics, but also evolutionary biology, around the concept of 'homology'. Indeed, it is related to a prevailing notion throughout the biological sciences in which certain instances of similarity are understood as homology (the latter indicating descent from a common ancestor). Although this idea first took hold in evolutionary biology, where taxonomies were constructed based on the theory that similar physical morphologies among species were likely to indicate common descent and relatedness, 20th century genetics and later bioinformatics have translated this idea of common descent to observations of similarity in genome sequences (Fujimura, 1999).

The population geneticists we interviewed, who study human samples collected primarily in Europe and North America, believe that individuals with similar SNP variation scores (interpreted as similar SNP patterns) are likely to have 'shared ancestry'. This is in part because of ideas of homology, but also in part because of research and theories from the fields of archaeology, physical anthropology, linguistic anthropology, and social anthropology. The population geneticists we interviewed accept the theory that modern humans originated relatively recently in Africa and then spread through different migratory patterns across the globe. Although there are varying ideas about the degree of difference among populations living on separate continents, the general belief expressed by our respondents is that the differences are recent in terms of hominid history, and some further assume that they are useful for distinguishing certain combinations of markers as originating from groups on certain continents.

Indeed, some population genetics researchers sometimes conceptualize ancestral and contemporary populations in terms of the major continents, for example, 'Africans' and 'Europeans'. In response, some (for example, Fullwiley, 2008; Marks, 1995; Ossorio and Duster, 2005) have argued that notions of continental ancestry do not always correlate with individuals' self-identified race, and reinvigorate ideas about the 'major races', as conceptualized in the 18th century by anthropologists (for example, Blumenbach, 1795). Further, as noted above, assessments of relationships between contemporary people and ancestral populations are layered with uncertainties, due in part to the lack of physical and genetic material with which to evaluate 'ancestors'.

The critique of continental ancestry applies appropriately, in our estimation, to admixture mapping genetic studies, the forerunner to GWAS, and we raise this issue here to distinguish the two (Fujimura et al., 2010). In admixture mapping genetic studies, researchers use groups they describe as 'recently admixed (for example, African-Americans) as their study populations. They typically identify these groups as having a mix of at least two different continental origins. These studies sometimes use language that assigns ancestries to chromosomes and DNA, such as 'chromosomes of entirely European or African origin' (Freedman et al., 2006). Admixture studies posit that particular 'chunks of DNA', or portions of chromosomes, have their 'ancestral origins' on particular continents.

Geographic descriptors became integral to how admixed populations were initially defined, because population genetic theories use geographic location to discuss gene flow between groups and designate ancestry differences. Among some GWAS researchers, the move from '*similar SNP variation scores*' to '*related genetic histories*' to '*shared ancestry*' is similarly accomplished through notions of geography. The researchers use algorithmic technologies to produce SNP similarities and differences, but the inferences of shared ancestry from SNP similarities are based on population genetic theories of human migration pathways around the world, which themselves are constructed via histories written through archaeology, physical anthropology, linguistic anthropology, and social anthropology. We call these practices of using geographical assessments to define genetic ancestry within the research designs and the technologies used to construct genetic populations '*genome geography*'.

Among our respondents, geography (that is, a physical location describing the 'origin' or current residence of the sample donor) is sometimes conceptualized in terms of national labels, regions of a continent, or smaller localities such as cities and towns. Geography is assessed sometimes through 'genetics' (that is, through SNP similarity), and sometimes through self-report. For example, the heavily used Hapmap samples and data are geographically labeled, which helps to explain why many of the researchers we interviewed attributed geographic meaning to differences in SNP frequencies when discussing the results of the Eigensoft output and why they interpreted these geographic differences to mean different ancestries. Thus, geographical considerations are used in producing what they consider to be ancestrally related individuals. For example, one researcher discussed the 'correlation' between the genetic clusters and geographical areas where study subjects were ascertained[20] including nation states and sites within nations.

> … it was only by typing hundreds of thousands of markers that you could actually even start to pull people apart into groups. But when you did … it did correlate with sort of Sweden versus

> Finland. And even within areas of Finland, depending on where patients had been ascertained … it correlated with the genetics as well.

In some of the GWAS conducted by the scientists we studied, geographical ancestry is assessed through methods other than SNP genotyping. For example, some researchers have tried to incorporate genealogical information based on where the individual's parents and grandparents were from.[21] Thus, genealogy, geographical location, and SNP genetics become inextricably tied together.

Though the concept of ancestral 'origins' is not at the forefront of some of the current biomedical GWAS discourse, it is prominent in related population genetic and evolutionary studies of variation and migration in humans. Here, designations of geographic origin go beyond continents in the evaluative assumptions attached to EIGENSTRAT output, as demonstrated by several recent studies of European populations which claim to be able to distinguish populations regionally, along northwest and southeast axes of 'European ancestry', and sometimes even along national boundaries.

Part of the reason for making such fine distinctions is that, in contrast to admixture mapping studies of US-based groups, which weave a tight relationship between genetic ancestry and continental race, the developers of Eigensoft view the level of ancestry relevant to their work as more clinal. 'Cline' is a concept developed in evolutionary biology to describe continuous rather than discretely bounded variations in traits among groups. A cline is 'a continuous gradation over space in the form or frequency of a trait' (Livingstone, 1962).

> In data sets with ancestry differences between samples, axes of variation often have a geographic interpretation: for example, an axis describing a northwest–southeast cline in Europe would have values that gradually range from positive for samples from northwest Europe, to near zero in central Europe, to negative in southeast Europe. (Price et al., 2006)

Still, despite the efforts of some GWAS researchers to avoid the use of race categories through EIGENSTRAT, the program carries with it a notion of genetic ancestry that relies on some of the same geographic ideas that circulate in admixture mapping studies.[22] Admixture mapping researchers studying US-based groups read particular race categories as admixtures of DNA sequences which they regard as 'originating' from different continents. For example, some of our respondents who conduct admixture mapping studies on US-based groups view African-Americans as carrying 'chunks of DNA' inherited from African slaves and European slave owners. They regard Latinos in the US as carrying DNA in their genomes inherited from indigenous Native Americans, Spanish and other European colonizers, and sometimes African slaves. In part, because such admixture studies begin with populations labeled 'Latino' and 'African-American', admixture mapping researchers have been viewed as using race or ethnic categories.[23] Despite the fact that GWAS use different technologies to produce and study populations, some of the researchers read similar stories of origins and migration into the SNP variation patterns of their research subjects. Given that some GWAS researchers use labels such as 'European' in their work, and given that the population geneticists who developed the Eigensoft package also helped develop admixture mapping methods, it is

unsurprising that GWAS research is sometimes interpreted as using socio-cultural categories of race or ethnicity.

Discussion: Genetic ancestry, genome geography, and race

This article has examined current research on the genetics of disease susceptibility in the context of debates about the use of race categories in such research. We have focused here on the practices and processes of GWAS in great detail to show how some geneticists are constructing new ways to measure 'genetic ancestry' differences, and we analyze how this emerging concept of ancestry does and does not relate to notions of race.

Some medical geneticists conducting GWAS have attempted to search for genetic markers for disease by developing a technology called EIGENSTRAT to help them to avoid the use of race categories. They affirmed that their genetic analyses 'were *not* about race' – whether defined by ideas about phenotypic characteristics such as skin color or self-described categories. They have sometimes worked with samples that were originally collected in terms of race or ethnic categories, but they did not use these categories in their analysis of correlations between SNP variation and disease. Some of them believe that race is an incorrect concept to use in disease genetics, because race categories are insufficient proxies for specifying groups in which individuals have relatively similar genetics. Instead, they use their EIGENSTRAT technology to produce numerical SNP variation scores that are similar for people with similar genetic marker patterns. In general, the geneticists we studied who are using the EIGENSTRAT technology praised its technical merits for aiding the search for genetic factors that may contribute to CCD.

Nevertheless, we have found in *some* GWAS articles, as well as in interviews with *some* GWAS researchers, a translation or slippage from clusters of genetically similar samples to categories of samples with similar 'genetic history' or 'shared ancestry'. Their use of 'shared ancestry' to describe the clusters that can be generated by the software points to particular assumptions and practices that are part of the production and circulation of the Eigensoft package.

Although 'ancestry' is not equivalent to race, 'shared ancestry' could be interpreted as race (as discussed below), especially when genetic ancestry is traced back to the major continents of Africa, Asia, Europe, and the Americas. It is these continental geographies that lend GWAS data to possible racial interpretations, that is, to the entangling of ancestry with race.

As we have shown, researchers use notions of geography to define and interpret ancestry in their research designs and in the technologies used to denote populations. We call this process of tracing genetics via 'ancestry' to geographic locations *genome geography*. Genome geography is deeply informed by scientific and socio-cultural discourses on human geography and human history, including theories of the origins and migrations of human groups.[24] Genotyped SNPs, Eigensoft renderings of those SNPs, and socio-cultural discourses of history and geography are tightly entangled through these socio-material practices.

Through these practices, these scientists produce simultaneously *different kinds of populations* and *population differences.* That is, concepts of population and population differences are part of both the production and outcomes of these studies. This article

examined the processes and concepts used to produce populations and population differences, and we shall now discuss some *consequences* of such productions of differences.

Downstream practices: Lost in translation

As noted, biomedical GWAS research need not denote populations with genetic ancestry labels when researchers use EIGENSTRAT. Although some of the researchers we studied insisted on *not* labeling populations, other GWAS studies have published population ancestry labels. The distribution and diversity of human genetic variation is a complex topic studied by researchers from diverse disciplines whose work is read by multiple audiences with differing views. In this context, the subtlety of the difference between race and ancestry may get *lost in translation*. Despite efforts to avoid the use of race categories, GWAS has the potential to be interpreted as focusing on race because some publications in the field link SNP variation clusters to 'ancestral' geographic locations or 'shared ancestries'. Thus, even though some researchers believe that race categories are sociohistorical concepts and that race is an incorrect concept for use in genetics, the notion of shared ancestry is often read as race by the media, the public, or other researchers.

For example, the geneticists we interviewed emphasized the point made earlier that GWAS research examines differences in the *frequency*, and not the existence per se, of disease-associated variants in different groups. This difference between frequency and existence is often mistakenly elided, overlooked, or even misunderstood as race-specific markers by the media and the public. In 2007, *New York Times* science writer Nicholas Wade read 'race' and race-specificity of marker variants into the results of one of the first GWAS on CCD, conducted by the Wellcome Trust on several thousand British patients with 'European ancestry'. While one could fault Wade for misreading the study, some geneticists admit that their publications and their universities' press releases do not make clear to those outside the field that the finding of a particular marker variant in a population does not preclude its existence in other populations, although these populations may differ in the frequency of its appearance. Some geneticists said that colleagues know what they mean by their terms, so they sometimes use shorthand terms that can be misread by people who are not in the field. Researchers also told us that their communications offices often use simplifying, provocative terms in press releases about their research results.[25] Because of this, their communiqués are sometimes interpreted as indicating that only certain groups carry disease alleles.

In another example, a medical genomics study conducted by researchers in Japan used Eigensoft to examine population substructure among Japanese citizens. They concluded that their samples (taken from people living in different areas of Japan and who self-identify as 'Japanese') cluster into groups that differ genetically. They argued for two genetic clusters of individuals, one cluster of Okinawans who were most likely descendants of ancient peoples such as the Jomon from the Ryukyu Islands, and a larger cluster corresponding to the descendants of ancient Yayoi (with possible Jomon mixture) from the main islands of Japan excluding the Ryukyu Islands. The researchers named a subset of the Yayoi cluster to be more closely related to 'Han Chinese'. They used archaeological theories about early settlers of what is now called Japan, and information on the location of the sampled individuals, to develop their arguments. Their study

concluded that the design of future GWAS for disease-associated SNPs must take into account genetic differentiation within the Japanese population, especially if there are differences in disease prevalence among geographical regions of Japan (Yamaguchi-Kabata et al., 2008). This is an interesting conclusion and rationale for conducting such a study, especially given the practices of many genomicists outside Japan who treat Japanese samples as coming from a roughly homogeneous population, and even 'lump' them with other samples in the category 'Asian'. Given that 'Asian' is often read as a race category, the articulation of sub-categories of Asians may be considered an advance on 'lumping'. However, these geneticists are producing new populations and population differences, linking genetics to 'ancestral' geographic locations using as their reference point some of the prevailing stories of human geography and migration to and within Japan. Although genome geography might be considered an advance over the five continental race categories defined by early anthropologists, it can also be read as producing new, albeit finer, genetically bounded categorical distinctions among peoples. Our research suggests that, contrary to emphasizing the notion that humans are all related, some studies of population substructure using GWAS statistics and methods are performing a kind of ancestry tracing, and finding differences that are often read as racial or ethnic differences.[26]

Categorical alternatives, choices, and scientific change: Implications for policy

As we have shown, contemporary statistical biomedical genetics is buttressed by a logic of difference. For example, the researchers we have studied focus on differences in genotypes and genetic susceptibilities between individuals and groups as explanations for differential disease risk and occurrence. Nevertheless, there are different ways to construct, represent, and operationalize difference in genetic variation studies. GWAS and EIGENSTRAT technologies provide medical geneticists with tools to analyze their samples *without* using *race* or *ethnic* categories. That is, they provide alternative ways of doing genetic research into complex diseases. Even when their samples have been collected using race or ethnic categories, they need not be analyzed in those categorical terms. Further, when using EIGENSTRAT to search for disease-associated SNPs, researchers need not use *ancestry* labels to describe their research samples; thus those who choose to label and publish their analyses in terms of ancestry categories are making a choice. They could instead choose to use EIGENSTRAT or other measures of SNP variation to conduct the regression analysis between cases and controls without ever framing their research in terms of ancestry clusters of SNP variation scores, and certainly without publishing plots of SNP variation scores that demarcate clusters with ancestry labels.

More generally, our study shows that biomedical genomic researchers have degrees of flexibility in their research designs and can avoid using race categories. They certainly are not required to use categories of race or ethnicity in their data *analysis*. The National Institutes of Health (NIH) Revitalization Act encourages the inclusion of women and racial minorities in clinical research and mandates that practitioners in clinical and basic biomedical research receiving federal funding should *report* on the diversity of their research subjects according to racial and ethnic categories designated by the Office of Management and Budget (OMB). However, with respect to biomedical

genomics research,[27] the regulations do *not require* either that the researchers must include members of all groups in their studies or that the researchers must *analyze* their samples using race categories. Indeed, the researchers we studied argued that the use of OMB categories in the actual *framing and analysis* of biomedical genomics research practices is *incorrect science*.

Nevertheless, others have reported on the pervasiveness of the OMB categories in many biomedical study designs (Epstein, 2007; Kahn, 2006). It may be that the policy has been interpreted by some genetics researchers as an edict, or even as an opportunity, to use race categories in their research. Our study, however, demonstrates the opposite response among other geneticists.[28]

Based on this example, we argue that, with caution and care, GWAS technologies can provide alternative means to conduct searches for genetic markers associated with complex diseases, without relying on race or ethnic categories. We further argue that these and related technologies could potentially contribute to transforming our institutional and cultural forms and frames regarding genetics and human disease, through the invention and use of categories of genetic similarities without attached labels or inferences of shared ancestry.

However, although the invention of a way of conducting human genetics research without using race categories provides an *opportunity*, its potential for transforming genetics research may be limited by downstream translation practices (as discussed above), the ways some researchers have subsequently used the Eigensoft software, the tenacity of concepts of race, and the many historical examples of seemingly neutral terms eventually coming to gain evaluative meanings.

Conclusion

Our study analyzed how scientists produce simultaneously *different kinds of populations and population differences*, sometimes by appealing to popular categories of race, ethnicity, or nationality, and sometimes to 'genetic ancestry'. We have explored *different differences* both in constructed populations as well as among the scientists who do the constructing. We have shown that the invention of new genetic concepts of ancestry still relies on old discourses, but also incorporates new knowledges, technologies, infrastructures, and political and scientific commitments. That is, our study revealed dynamic and fluid socio-technical 'assemblages'[29] of actors in various disciplines responding to each other and to sociopolitical contexts, making decisions about how to do the research based on their assessments of scientific accuracy, new technologies, disciplinary commitments, political expediency, and personal politics. Each assemblage also includes particular ideas and discourses of history and geography as they become entangled with research on disease. These actors, discourses, knowledges, infrastructures, and commitments are dynamically inter-related and sedimented, thus defying rapid change in ways that resemble the obduracy of bureaucratic structures. Although we have identified new actors – both human and technical – working to *avoid* the use of race categories in biomedical genetics research, it is not clear how much they can change the institutionalized and historical practices within the larger assemblage.[30]

In this case, the medical geneticists we studied are aware that EIGENSTRAT and its output are constructed from the statistical assumptions that go into PCA, as well as from the assumptions adopted from other disciplines like population genetics. They take EIGENSTRAT results to be *better* representations that generate more informative analyses, but they also understand that its results are constructed. Neither they, nor we, naturalize genetic marker readouts, genotypes, and SNP similarity groups. The medical geneticists in our study used EIGENSTRAT in ways that they hoped would avoid the invocation of race, while some population geneticists studying groups in North America and Europe used EIGENSTRAT to develop clusters that they labeled in ways that led other researchers, members of the media, and members of the general public to read race and ancestry categories into their studies.[31] Although genetic ancestry is not race, these ancestry categories are often read as race and ethnicity, especially when the ancestry labels match national, tribal, and other designations that have socio-cultural meaning and carry political implications.

Based on this and other STS case studies, we argue more broadly that any technology is developed within an assemblage that has its own genesis, history, and outcome. STS research on theory-methods packages (Fujimura, 1988), experimental systems (Fujimura, 1996, Jordan and Lynch, 1992, Rheinberger, 1997), platforms (Keating and Cambrosio, 2003), and socio-technical assemblages (Callon and Rabeharisoa 2008; de Laet and Mol, 2000) has shown that technologies constructed for one purpose can be adopted into new sites and used in different, sometimes unintended, ways. In this case, changing researchers' use of race categories to the use of SNP variation scores in other sites in the assemblage or in other assemblages may be difficult if race is sedimented within the larger assemblage. To make SNP variation scores, rather than race, the tool for correcting for stratification in GWAS, researchers who support this shift will need to take upon themselves the task of convincing their colleagues (particularly those who insist on bringing race back in when unwarranted by the data) that their uses of EIGENSTRAT are just *better science*.

Acknowledgments

We thank the researchers who participated in our study, for their willingness to discuss these issues with us, and their generosity in allowing us to interview them and observe their research processes. We are grateful to Kjell Doksum, Pilar Ossorio, Tanya Cook, Cabell Gathman, and Hanna Grol-Prokopczyk for their work on our larger project. This article benefited from extensive comments from Deborah Bolnick, Lundy Braun, Jane Calvert, Alberto Cambrosio, John Dupre, Troy Duster, David Jones, Jay Kaufman, John Novembre, Maureen O'Malley, Diane Paul, Vololona Rabeharisoa, Pamela Sankar, Elliott Sober, Kim Tallbear, Sharon Traweek, Claire Wendland, and Alice Wexler, and from comments received at numerous presentations. We are also grateful to editor Michael Lynch and three anonymous referees.

Funding

This work was supported by NSF [grant number 0621022], NIH [grant number R03HG005030-02], and grants from the University of Wisconsin Institute for Race and Ethnicity and the University of Wisconsin Graduate School.

Notes

1. Braun, 2006; Cooper et al., 2003; Fausto-Sterling, 2004; Fujimura et al., 2008; Fullwiley, 2008; Kahn, 2006; Lee, 2003; Marks, 1995; Montoya, 2007; Ossorio and Duster, 2005; Sankar, 2008; Shields et al., 2005; Wailoo and Pemberton, 2006.
2. For socio-material practices of knowledge production, see Suchman (2002) and Fujimura (2006). For related concepts, see Haraway's (1991) material-semiotic practices and Barad's (1998) notion of intra-action.
3. But see Bolnick (2008) for a critique of two publications that used the *Structure* program and its algorithms to generate clusters of samples that seem to correspond to continental races. Bolnick points to the assumptions underlying the probability equations built into *Structure*, that warrant consideration when drawing conclusions about groups. She critiques the way the two human genetic variation studies and their audiences (mis)interpreted the results of the program as evidence of race, rather than as evidence of clinal variation, in the samples studied.
4. These include the development of faster and cheaper genotyping technologies, the Human Haplotype Maps, annotated databases of human genome sequences, and statistical genetics techniques.
5. Some researchers argue that the larger part of common complex disease causation is based in environmental exposures and social inequalities and that research and therapeutic resources should be allocated to these causal factors instead of to the investigation of genetic causes (for example, Chaufan, 2007; Krieger and Fee, 1994). While geneticists we have interviewed agree that genetic causes are not predominant, they maintain that exploring such causes is important to furthering our overall understanding of disease etiology. Indeed, geneticists have been discussing, at least amongst themselves, the limitations of GWAS for explaining complex disease phenotypes for some time now. In 2007, some of our respondents stated that GWAS findings would only account for a small fraction of the etiology of the diseases they were studying; nevertheless, they were committed to their studies for various reasons. More recently, the *New England Journal of Medicine* published discussion pieces on both the unrealized promises (Goldstein, 2009; Kraft and Hunter, 2009) and contributions (Hirschhorn, 2009) of GWAS methods.
6. Indeed, on the clinical side, Rabeharisoa and Bourret (2009) found that the cancer geneticists and psychiatric geneticists they studied tended to use information about genetic mutations related to diseases as only one aspect of a multi-factorial representation of disease, rather than a causal representation. They found that genetics did not trump other potential sources of explanation for heterogeneous pathologies encountered in the clinics they studied.
7. SNP markers are assayed by genotyping methods, which determine the SNP allele, or variant, at each SNP marker in each individual sample. Two of the most common 'platforms' for genotyping are Affymetrix microarrays, which distinguish single nucleotide variants based on hybridization efficiency of the sample to a complementary sequence immobilized on the array, and Illumina beadchips, which use allele-specific single-base extension of sample sequence, after hybridization to probe sequences on beads arrayed on the chip.
8. What constitutes 'enough' individuals and markers is a decision that researchers make using statistical calculations.
9. The technologies used in genome-wide association studies are currently changing, and soon disease association studies will be using whole genome sequences rather than a set of genotyped SNPs. This article focuses on recent research based on technologies that use a set of genotyped SNPs.
10. Initially, researchers focused on a subset of these 10 million SNPs, trying to catalog those that appear in many sampled individuals and groups, which they call 'common' variants.

However, they are now turning their attention to 'rare' SNPs and copy number variants that may have more severe effects and thus play a larger role in complex diseases than common variants (Hardy and Singleton, 2009; McClellan and King, 2010).

11. To clarify, this distinction between signal and noise is one made by the researchers we studied. Our position is that what counts as 'signal' and 'noise' depends entirely on the problem researchers pose and how they formulate their research designs. This applies to all researchers, including ourselves.
12. Researchers who identify as medical geneticists typically conduct molecular genetic experiments on medically related topics. Many have a clinical or medical background. In our GWAS research sites, some of the medical geneticists we studied had medical training, and some also had statistical training. Those who identify as population geneticists typically have a background in evolutionary biology and genetics, as well as training in mathematics, applied computer science, and/or statistics.
13. The statistics behind the software program are designed specifically for large-scale studies – those that test 100,000 or more SNPs in many thousands of people. EIGENSTRAT is thought to be less effective for finding significant disease-associated SNPs in smaller, more targeted studies involving fewer genetic variants.
14. Our choice to study EIGENSTRAT and the larger Eigensoft package was based on our discovery during ethnographic research that GWAS researchers had adopted this recently designed technology in their search for medically significant SNPs. We were not aware of EIGENSTRAT before engaging in ethnographic fieldwork. However, our choice of researchers to study was based on their efforts to find ways to avoid using categories of race in their genetics research.
15. This is an important qualifier. We are writing here only about the medical geneticists and population geneticists we have studied.
16. Accounting for the 'genetic history' of samples is in their view the same as 'adjusting for population substructure' or 'adjusting for population stratification'.
17. The EIGENSTRAT algorithm does not encode ideas of human evolution or try to assign individuals to groups, as do other population genetic software programs like 'Structure'. In Structure, because the user has to specify the *number* of clusters or groups in advance, the user's 'accepted wisdom' regarding human evolution indirectly plays a key role because it informs their choice of number of clusters to request of the program. Thus, it influences the resulting assignment of individuals to the 'best' number of groups.
18. The medical geneticists we studied are currently looking at common variation, which means that the variants of each SNP will appear in many populations, although at different frequencies. Generally, it would be difficult to say that a SNP variant is specific to one group or ancestry, because one would have to test everyone else to say that it appears exclusively in that group. However, some admixture mapping genetic studies assume that certain blocks of SNPs do have a discernible ancestry.
19. We can only mention these briefly here.
20. Ascertainment sites are the clinics where patients and their controls were recruited into a study.
21. Even where detailed records going back many generations are available, GWAS tend to include information only as far back as grandparents. We assume this is because of the desire for comparability of data across sites and samples; many people are unable to give information beyond their grandparents' generation.
22. Despite the extensive human, technological, computational, and financial resources required, GWAS have now overtaken admixture studies, although some geneticists still conduct the latter because of resource constraints.

23. 'Hispanic' and 'Latino/a' groupings emerged from a brutal history of colonization that some geneticists now reiterate through genetics. Montoya (2007) describes this history told through genetics as a second colonization which he calls 'bioethnic conscription'. We have found that some geneticists understand that their genetic research is an addition – and not the final word – to histories already available from the people whose DNA they study; the anthropologists who study the histories, languages, and cultures of those same people; and the historians who study their oral histories and documents where available. Other geneticists appear to believe more strongly in their publications that genetics is the final word.
24. As social studies of science have shown, the boundaries between what count as scientific and what count as socio-cultural discourses are often fuzzy.
25. Slippage from frequency statements to race statements is sometimes made by researchers themselves before translation from research to readers. However, we are writing here primarily about the geneticists we interviewed who made concerted efforts not to make this slippage. We also note here that university communications offices often facilitate the slippage in part because research on race and genetics is provocative and therefore 'sells'. This marketing strategy implicates both sellers and their audiences.
26. Japan is an ethnically stratified society where those called 'Okinawan' and 'Chinese' often suffer discrimination. Injecting constructions of genetic differences that map onto these folk terms may have consequences that such studies may not intend.
27. The guidelines implemented in accordance with section 492B of the Public Health Service Act, added by the NIH Revitalization Act of 1993, Public Law (PL) 103-43 state that: 'It is the policy of NIH that women and members of minority groups and their subpopulations must be included in all NIH-funded clinical research, unless a clear and compelling rationale and justification establishes that inclusion is inappropriate with respect to the health of the subjects or the purpose of the research' (National Institutes of Health, 2001). Note that this refers to *clinical* research working directly with human subjects and studying for example their responses to treatments. In GWAS, blood or other biomaterials are collected from human subjects and anonymized, and genotyping and analysis of DNA samples often happens elsewhere and does not involve interaction with research subjects. In some cases, researchers have been explicitly told to include samples from only 'one single ethnic group', and not from other 'ethnic groups', because of concerns about confounding due to population substructure.
28. This may have been partially a response to critiques by social scientists of the use of race categories in genetic research with respect to the Human Genome Diversity Project and the National Human Genome Research Institute (NHGRI)'s International Haplotype Mapping Project (HapMap). Some of the geneticists we interviewed had been involved in the HapMap project and were already sensitive and sympathetic to these critiques. After initial controversies over the HapMap project, NHGRI established a special working group involving both social scientists and geneticists to deliberate on the use of race in genetics research.
29. The notion of assemblages was originally developed by Deleuze and Guattari (1987).
30. Fujimura (1988) showed how theory-methods packages were used to radically transform cancer and other biological and biomedical research enterprises.
31. Disciplinary differences may partly influence these different tendencies, but there are medical geneticists who, like human population geneticists, discuss genetic differences and attribute ancestry to such differences. Nevertheless, there are disciplinary differences that made for sometimes tense interactions in the collaborations between the medical

geneticists and population geneticists around EIGENSTRAT and around GWAS. We are writing a separate article on the complexities of interdisciplinary collaborations in large-scale genomics.

References

American Society for Human Genetics (2008) *The American Society of Human Genetics Ancestry Testing Statement*. Bethesda, MD: American Society for Human Genetics. Available at: www.ashg.org/pdf/ASHGAncestryTestingStatement_FINAL.pdf (accessed 22 August 2010).

Barad K (1998) Getting real: Technoscientific practices and the materialization of reality. *Differences* 10(2): 87–128.

Blumenbach JF (1795) *De generis humani varietate nativa liber* [*On the Natural Variety of Mankind*]. Goettingae: Vandenhoek et Ruprecht.

Bolnick DA (2008) Individual ancestry inference and the reification of race as a biological phenomenon. In: Koenig B, Lee S and Richardson S (eds) *Revisiting Race in a Genomic Age*. New Brunswick, NJ: Rutgers University Press, 77–85.

Braun L (2006) Reifying human difference: The debate on genetics, race, and health. *International Journal of Health Services* 36(3): 557–573.

Brookes AJ (1999) The essence of SNPs. *Gene* 234(2) (8 July): 177–186.

Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, et al. (2003) The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine* 348(12) (20 March): 1170–1175.

Callon M and Rabeharisoa V (2008) The growing engagement of emergent concerned groups in political and economic life: Lessons from the French Association of Neuromuscular Disease Patients. *Science, Technology, & Human Values* 33(2): 230–261.

Chaufan C (2007) How much can a large population study on genes, environments, their interactions and common diseases contribute to the health of the American people? *Social Science & Medicine* 65(8): 1730–1741.

Cooper R, Kaufman JS, and Ward R (2003) Race and genomics. *New England Journal of Medicine* 348(12) (20 March): 1166–1170.

De Laet M and Mol A (2000) The Zimbabwe bush pump: Mechanics of a fluid technology. *Social Studies of Science* 30(2): 225–263.

Deleuze G and Guattari F (1987) *A Thousand Plateaus*: *Capitalism and Schizophrenia*. Minneapolis: University of Minnesota Press.

Duster T (2003) Buried alive: The concept of race in science. In: Goodman A, Heath D, and Lindee M (eds) *Genetic Nature/Culture: Anthropology and Science Beyond the Two Culture Divide.* Berkeley: University of California Press, 258–277.

Epstein S (2007) *Inclusion: The Politics of Difference in Medical Research*. Chicago: University of Chicago Press.

Fausto-Sterling A (2004) Refashioning race: DNA and the politics of health care. *Differences: A Journal of Feminist Cultural Studies* 15(3): 1–37.

Freedman ML, Haiman CA, Patterson N, McDonald GL, Tandon A, Waliszewska A, et al. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proceedings of the National Academy of Sciences* 103(38) (19 September): 14068–14073.

Fujimura JH (1988) The molecular biological bandwagon in cancer research: Where social worlds meet. *Social Problems* 35(3): 261–283.

Fujimura JH (1996) *Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer*. Cambridge: Harvard University Press.

Fujimura JH (1999) The practices of producing meaning in bioinformatics. In: Fortun M and Mendelsohn E (eds) *The Practices of Human Genetics*. Dordrecht, the Netherlands: Kluwer Academic Publishers, 49–87.

Fujimura JH (2006) 'Sex genes': A critical socio-material approach to the politics and molecular genetics of sex determination. *Signs: Journal of Women in Culture and Society* 32(1): 49–82.

Fujimura JH, Duster T, and Rajagopalan R (2008) Race, genetics, and disease: Questions of evidence, matters of consequence. *Social Studies of Science* 38(5): 643–656.

Fujimura JH, Rajagopalan R, Ossorio PN, and Doksum K (2010) Race and ancestry: Operationalizing populations in human genetic variation studies. In: Jones D and Whitmarsh I (eds) *What's the Use of Race: Modern Governance and the Biology of Difference*. Cambridge, MA: MIT Press, 169–186.

Fullwiley D (2008) The biologistical construction of race: 'Admixture' technology and the new genetic medicine. *Social Studies of Science* 38(5): 695–735.

Goldstein DB (2009) Common genetic variation and human traits. *New England Journal of Medicine* 360(17) (23 April): 1696–1698.

Haraway DJ (1991) *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York: Routledge.

Hardy J and Singleton A (2009) Genomewide association studies and human disease. *New England Journal of Medicine* 360(17): 1759–1768.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* 106(23) (9 June): 9362–9367.

Hirschhorn JN (2009) Genome-wide association studies – illuminating biologic pathways. *New England Journal of Medicine* 360(17) (23 April): 1699–1701.

Jordan K and Lynch M (1992) The sociology of a genetic engineering technique: Ritual and rationality in the performance of the 'plasmid prep'. In: Clarke A and Fujimura JH (eds) *The Right Tools for the Job*. Princeton: Princeton University Press, 77–114.

Kahn J (2006) Genes, race, and population: Avoiding a collision of categories. *American Journal of Health Policy* 96(11): 1965–1970.

Keating P and Cambrosio A (2003) *Biomedical Platforms: Realigning the Normal and the Pathological in Late-twentieth-century Medicine*. Cambridge, MA: MIT Press.

Kraft P and Hunter DJ (2009) Genetic risk prediction – are we there yet? *New England Journal of Medicine* 360(17) (23 April): 1701–1703.

Krieger N and Fee E (1994) Man-made medicine and women's health: The biopolitics of sex/gender and race/ethnicity. *International Journal of Health Services* 24(2): 265–283.

Lee SS (2003) Race, distributive justice and the promise of pharmacogenomics: Ethical considerations. *American Journal of Pharmacogenomics* 3(6): 385–392.

Livingstone F (1962) On the non-existence of human races. *Current Anthropology* 3: 279–281.

McClellan J and King M-C (2010) Genetic heterogeneity in human disease. *Cell* 141(2): 210–217.

Marks J (1995) *Human Biodiversity: Genes, Race, and History*. New York: Aldine de Gruyter.

Montoya M (2007) Bioethnic conscription: Genes, race and Mexicana/o ethnicity in diabetes research. *Cultural Anthropology* 22(1): 94–128.

National Institutes of Health (2001) NIH Policy and Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research – Amended, October, 2001. Available at: http://grants.nih.gov/grants/funding/women_min/guidelines_amended_10_2001.htm (accessed 22 August 2010).

Ossorio P and Duster T (2005) Race and genetics: Controversies in biomedical, behavioral, and forensic sciences. *American Psychologist* 60(1): 115–128.

Patterson N, Price AL, and Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* 2(12): e190.

Pennisi E (2007) Breakthrough of the year: Human genetic variation. *Science* 318(5858) (21 December): 1842–1843.

Price AL, Butler J, Patterson NJ, Capelli C, Pascali VL, Scarnicci F, et al. (2008) Discerning the ancestry of European Americans in genetic association studies. *PLoS Genetics* 4(1): e236.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8): 904–909.

Rabeharisoa V and Bourret P (2009) Staging and weighting evidence in biomedicine: Comparing clinical practices in cancer genetics and psychiatric genetics. *Social Studies of Science* 39(5): 691–715.

Rebbeck TR and Sankar P (2005) Ethnicity, ancestry, and race in molecular epidemiologic research. *Cancer Epidemiology Biomarkers & Prevention* 14(11): 2467–2471.

Rheinberger HJ (1997) *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford, CA: Stanford University Press.

Sankar P (2008) Moving beyond the two-race mantra. In: Koenig BA, Lee SS-J, and Richardson SS (eds) *Revisiting Race in a Genomic Age.* New Brunswick, NJ: Rutgers University Press, 271–284.

Shields A, Fortun M, Hammonds EM, King PA, Lerman C, Rapp R, and Sullivan PF (2005) The use of race variables in genetic studies of complex traits and the goal of reducing health disparities. *American Psychologist* 60(1): 77–103.

Suchman L (2002) Located accountabilities in technology production. *Scandinavian Journal of Information Systems* 14(2): 91–105.

Suchman L (2007) *Human-Machine Reconfigurations: Plans and Situated Actions*, 2nd edn. Cambridge, UK: Cambridge University Press.

Wade N (2007) Researchers detect variations in DNA that underlie seven common diseases. *New York Times* (7 June): 32.

Wailoo K and Pemberton SG (2006) *The Troubled Dream of Genetic Medicine: Ethnicity and Innovation in Tay–Sachs, Cystic Fibrosis, and Sickle Cell Disease*. Baltimore: Johns Hopkins University Press.

Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, et al. (2008) Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: Effects on population-based association studies. *American Journal of Human Genetics* 83(4): 445–456.

Biographical notes

Joan Hideko Fujimura is Professor of Sociology at the University of Wisconsin-Madison and Professor of Science and Technology Studies in the Robert F. and Jean E. Holtz Center. She has written on the sociology of genetics, molecular biology, biotechnology, biomedicine, HIV-AIDS research, and systems biology. Other recent publications include 'Postgenomic Futures', *New Genetics and Society* (August 2005), 'The Science and Business of Genetic Ancestry Testing', *Science* (19 October 2007), and 'Calculating Life? A Sociological Perspective on Systems Biology', *EMBO Reports* (August 2009). In addition to the research discussed in this article, Fujimura is also studying interdisciplinarity, collaboration, and 'big science' in systems biology and epigenetics programs. She is also continuing her comparative analysis of Japanese and US genomics.

Ramya Rajagopalan is a postdoctoral research associate in the Department of Sociology and the Holtz Center for Science and Technology Studies at the University of Wisconsin, Madison. Rajagopalan received her doctoral degree in genomics and molecular biology from the Massachusetts Institute of Technology in 2007. She has coauthored (with editors Joan H. Fujimura and Troy Duster) the introduction to a special issue on race, genomics, and medicine in *Social Studies of Science* (October 2008). Along with Fujimura, she is currently using ethnographic methods in two projects that examine the life sciences, one around the use of concepts of population in human genetic variation research, and one around interdisciplinarity, work organization and collaboration in systems biology and epigenetics research.