JOAN H. FUJIMURA AND RAMYA M. RAJAGOPALAN*

Race, Ethnicity, Ancestry, and Genomics in Hawai'i: Discourses and Practices

ABSTRACT

This paper examines how populations in a multiethnic cohort project used to study environmental causes of cancer in Hawai'i have been reorganized in ways that have contributed to the racialization of the human genome. We examine the development of two central genomic data infrastructures, the multiethnic cohort (MEC) and a collection of reference DNA called the HapMap. The MEC *study populations* were initially designed to examine differences in nutrition as risk factors for disease, and then were repurposed to search for potential genomic risk factors for disease. The biomaterials collected from these populations became institutionalized in a data repository that later became a major source of "diverse" DNA for other studies of genomic risk factors for disease. We examine what happened when the MEC biorepository and dataset, organized by ethnic labels, came to be used, in conjunction with the data from the HapMap *reference* populations, to construct human population genetic categories.

Developing theory on genomic racialization, we examine (1) how and why Hawai'i became sited as a "virtual natural laboratory" for collecting and examining biomaterials from different ethnic groups, and the consequences of the transformation of those local Hawaiian ethnic groups into five racial and ethnic OMB categories meant to represent global continental groups for genomic studies. We then discuss (2) how

*Fujimura: University of Wisconsin–Madison, Department of Sociology, 1180 Observatory Drive, Madison, WI 53706–1393,fujimura@ssc.wisc.edu. Rajagopalan: University of California, San Diego, Institute for Practical Ethics, 9500 Gilman Drive, MC 0406, La Jolla, CA 92093, rmrajagopalan@ucsd.edu.

The following abbreviations are used: ABC, unnamed cancer research institute in Hawai'i; CCD, common complex disease; DNA, deoxyribonucleic acid; GWAS, genome-wide association study; HapMap, haplotype map; HGP, Human Genome Project; MEC, multiethnic cohort; NCI, National Cancer Institute; NHGRI, National Human Genome Research Institute; NIH, National Institutes of Health; OMB, Office of Management and Budget; SNP, single nucleotide polymorphism; XYZ, unnamed cancer research institute in Los Angeles.

Historical Studies in the Natural Sciences, Vol. 50, Number 5, pps. 596–623. ISSN 1939-1811, electronic ISSN 1939-182X. DOI: https://doi.org/10.1525/hsns.2020.50.5.596.

this transformation, via the geneticists' effort to standardize the study of genomic risk for disease around the globe, led to the construction of humans as statistical genetic resources and entities for genomic biomedicine and the human population genetics discipline. Through this transformation of populations and biorepositories, we argue (3) that the twenty-first century has seen the intertwining of "race," "population," and "genome" via large-scale genomic association studies. We show *how* "race" has become imbricated in human population genetics and genomic biomedicine.

This essay is part of a special issue entitled *Pacific Biologies: How Humans Become Genetic*, edited by Warwick Anderson and M. Susan Lindee.

KEY WORDS: Hawai'i, genetic variation, race and ethnicity, infrastructures, human population genetics, genomic racialization

INTRODUCTION

The introduction of statistics into medical research has over the years contributed to medical knowledge and therapies. But the use of statistics in recent "big data" biomedical genomic research has brought a particular way of thinking about and framing human similarity and difference. This paper examines how the multiethnic cohort (MEC) study populations in epidemiology and cancer research have been repurposed for genomic studies that have intensified the use of racialized categories in genomic research in and about Hawai'i. To do so, it examines the decisions, practices, processes, and technological infrastructures that produced new methods for studying human genomic differences in the context of biomedical studies of genomic risks for common complex diseases (CCDs). These new methods have been used by some geneticists, sociologists, and media members to justify old ways of thinking "racially" about genetics. This paper shows *how*, in the US, old ideas of racial difference became embedded into new genomic biomedical research using "big data" and statistical methods.

Genetics and genomics are complicated, but so are notions of race. It is a commonly used concept in the US, where it is often used and understood differently depending on the user and the context. Sociologists and historians have studied how the concept has been developed and deployed for at least the past century. When the term "race" is not used, related terms such as "ethnicity," "ancestry," and "national origin" are used in its stead. Cultural, political, and social practices of classifying humans into various groups for political, economic, and sociocultural purposes have a long and contentious

history. "Race," as we employ the term, is a socially and historically constructed concept, not a set of biological categories.

We will examine how this socially, situationally, historically constructed concept and its related racialized categories have been used to develop populations for use in epidemiology and in statistical genetics research. But what is a population? There is no "natural definition" telling scientists what is and what is not a population. Rather, through the work processes they engage in, and the tools they craft and use, researchers decide who constitutes a population, and what defines a population's boundaries. It is through these work practices that "race" creeps into the production of population categories.

The work practices and tools we examine include the production of two central genomic data infrastructures used to construct genomic categories that, despite some researchers' efforts to the contrary, are routinely mobilized with racial and ethnic group labels. We begin by discussing how Hawai'i became viewed as a "natural laboratory" for studies of human genetic diversity and a site for developing the MEC and its attendant repository of biological samples and medical history data. We show how this Hawai'i–Los Angeles set of multiethic *study populations* (hereafter, the MEC study populations, or the MEC), initially constructed for examining ethnic differences in nutrition as potential risk factors for CCD, became repurposed to build a biological data repository of diverse DNA for use in case-control genome-wide association studies (GWAS) that sought genetic risk factors for CCD. Among these transformations, the MEC repository organized by ethnic group served as a template for constructing human population genetic categories. The second database infrastructure we examine is the International Haplotype Mapping (HapMap) Project—through which the *reference populations* were constructed—organized by the US National Institutes of Health's National Human Genome Research Institute (NIH, NHGRI) to collect DNA from around the globe, categorized by ethnic, city, national, and regional labels.

Through the repurposing of the MEC categories for genomic disease research, some of the original MEC populations became "study populations" or DNA samples from groups of people for use in the search for correlations between disease and human genomic variations in the form of single nucleotide polymorphisms (SNPs). Geneticists compared these study populations with HapMap "reference populations," or DNA samples from groups of people from whom human geneticists collected and cataloged SNPs before they correlated the SNPs with disease. Through this repurposing, using Donna Haraway's notion of dominant concepts defining different eras (race for the

early 1920s, population for the 1940s, and genome for the 1990s),[1] the twenty-first century has seen the intertwining of all three via GWAS. We show how social notions of race became embedded in both study and reference populations via GWAS, thereby entrenching racialized ideas in infrastructures for constructing human research populations and bringing back notions of race as innately biological and genetic.

In our analysis, we differentiate "genomic" from "genetic" in describing the kinds of categories of human groupings that researchers construct in their work. Though the terms are often used interchangeably by biomedical researchers and in popular discourse, there are many ways in which genome-era scientific practices and methods (since the early 2000s or so) differ from earlier twentieth-century laboratory practices in genetics. The rhetorical force of the term "genetic" to describe DNA-centric conclusions invokes stronger connotations of heredity, intergenerational passing on of traits, and biological durability than either genetics or genomics *research* has demonstrated. That is, many historical and contemporary uses of the term "genetic" have misleadingly and inappropriately lent stronger credence to these connotations than current data have provided for the majority of human traits, diseases, and behaviors subjected to the lens of genetic analysis. In tracking the work of genome scientists, we are expressly trying not to signal any of these connotations, because we have observed that genomic practices, logics, language, and associations have been used to capture what amount to more diffuse, tenuous, heterogeneous, and probabilistic relationships to disease, medicine, health, and intergenerational transmission.

Our methods include ethnographic research, interviews, and documentary analyses over several research sites and many years.

THE STUDY POPULATIONS: HAWAI'I AS A NATURAL LABORATORY

Building the Multiethnic Cohort (MEC) Biorepository for Prospective Studies of Diet and Disease

Native Hawaiians had been the primary residents of the Hawaiian Islands before armed forces aligned with American businessmen (sugar plantation

1. Donna Haraway, *Modest_Witness@_.__: Feminism and Technoscience* (New York: Routledge, 1997).

owners and descendants of missionaries), with the approval of the US government, overthrew the independent Kingdom of Hawai'i and its reigning monarch Queen Liliuokalani in 1893. These businessmen and their compatriot American missionary families colonized the islands and established plantations. When native Hawaiians resisted being put to work in their fields, these businessmen began to import contract laborers to the islands to work on sugar plantations. These laborers included, among others, individuals from Norway, Portugal, Puerto Rico, China, Korea, Japan, and the Philippines.[2]

Because of this ethnic diversity, Hawai'i has been treated by different generations of scientists as a "natural laboratory" in several eras. Before American colonization, although there were status differences, notions of race and racial differences had not existed in Hawai'i. But "by 1890, immediately before the United States' imperialist expansion into Cuba, Puerto Rico, the Philippines, Guam and Samoa, instructions to census takers followed a pattern already set—of compulsory racialization and simultaneously a struggle to figure out what to do with nonwhites. By the 1920s, not surprisingly, there was much more 'race talk' in Hawai'i. . . . "[3] In the 1920s and 1930s, Hawai'i became a natural laboratory site to study "racial science" for Australian physical anthropologists and Chicago School sociologists.[4] These researchers at first viewed "race mixing" as leading to the degeneration of the human species, and envisioned Hawai'i as a natural laboratory for studying so-called hybridization because of the high rates of intermarriage. Continental US eugenicists of that time period viewed miscegenation as a serious "problem" because of the increasing numbers of immigrants from the East. As Domínguez notes,

> the presumed purity and sameness of "the white race" was to be preserved statistically even, or perhaps especially, when there was ample evidence of sexual and familial mixing in various regions, states, and territories . . . "Any mixture of White and some other race" in the 1930 US census "was to be

2. Ronald Takaki, *Pau Hana: Plantation Life and Labor in Hawaii, 1835–1920* (Honolulu: University of Hawaii Press, 1983). Takaki shows that contract laborers were just another line item on purchasing lists that also included flour and wire.

3. Virginia R. Domínguez, *White by Definition* (New Brunswick, NJ: Rutgers University Press, 1986).

4. Warwick Anderson, "Hybridity, Race, and Science: The Voyage of the *Zaca*, 1934–1935." *Isis* 103, no. 2, (2012): 229–53; and Christine Leah Manganaro, "Assimilating Hawai'i: Racial Science in a Colonial 'Laboratory,' 1919–1939" (PhD dissertation, University of Minnesota, 2012).

reported according to the race of the parent who was not White; mixtures of colored races were to be listed according to the father's race, except Negro-Indian."[5]

In the end, ironically, these anthropologists and sociologists yielded rich data that pointed researchers to a view of human biology as dynamic and plastic, leading them to conclude that race mixing among people in Hawai'i appeared to be producing a reinvigoration of the species.[6] Chicago School sociologists like Robert Park, Andrew Lind, and Romanzo Adams also became hopeful that the high percentage of intermarriages and "mixed race" peoples in Hawai'i would ultimately produce a de-racialized society. Indeed, Adams "exaggerated the degree of interracial reproduction and suggested that the territorial population was well on its way to complete biological amalgamation."[7] However, their view of racial harmony was based on a biological amalgamation, that is, that harmony could only come with biological racial mixing.[8] From the 1920s through statehood (1959), social scientists also argued that race mixing with whites increased the fitness of Hawaiians and immigrant Asians to help prepare them for their roles as consumer capitalists and democratic citizens.[9] Theirs was a biological and cultural explanation for fitness as American "ideal citizens." And, in the end, this celebration of mixed-race identities was not reflected in mainstream continental US thinking, where typologies and racial classifications spread and took firmer root across the nation.[10]

Constructing Ethnic Categories to Study Diets as Risk Factors for Cancers

In the 1970s, Hawai'i was once again envisioned as a "natural laboratory." However, in contrast to the previous eras of social scientific studies, the new era began with the aim of understanding environmental risks of various cancers

5. Virginia R. Domínguez, "Exporting U.S. Concepts of Race: Are There Limits to the U.S. Model?," *Social Research* 65, no. 2 (1998): 369–99, on 384–05.

6. Anderson, "Hybridity, Race and Science" (ref. 4).

7. Manganaro, "Assimilating Hawai'i" (ref. 4).

8. An even more recent version of this idea of mixed race as leading to racial harmony was an article entitled "Want to be less racist? Move to Hawai'i," published by Moises Velasquez-Manoff in the *New York Times* on March 4, 2019. In response, the Popolo website hosted by Akiemi Glenn published the following statement: "Want to explore race in Hawai'i? Center those most impacted by it."

9. Manganaro, "Assimilating Hawai'i" (ref. 4).

10. Anderson, "Hybridity, Race and Science" (ref. 4).

based on cultural differences, primarily differences in diet, exercise activity, and other health measures described as "lifestyle factors." Researchers were not initially interested in innate genetic risks. Nevertheless, genomic differences organized by race entered the scene in the 1990s via technologies for studying human population genetic differences and collaborations with researchers interested in these differences.

Because Hawai'i was home to many recent immigrant groups, all located within a small geographic space, epidemiologists at a cancer research institute in Hawai'i (hereafter ABC) envisioned the islands as a "natural laboratory" for research on diet, nutrition, and disease. In this case, the epidemiologists specifically selected research subjects by their ethnic identities for inclusion in their study cohorts; that is, the epidemiologists viewed ethnic groups themselves as proxies for variations in diet and nutrition. Their diet and cancer studies began in the early 1980s, first as cross-sectional surveys in Hawai'i and migrant studies, and later as population-based case-control studies of various cancers including breast, prostate, lung, colorectal, stomach, endometrial, and thyroid. These studies included people from different ethnic groups to study the association between diets, nutrition, physical activity, and cancers.

> The initial studies we did are what are called case-control studies, which are less expensive to do and easier to do initially, and we were able to get funding and support to do case-control studies looking at breast cancer and colorectal cancer, the more common cancers. So everything evolved from there, because eventually, you do a sufficient number of case-control studies, which raises certain issues and questions, that you'd like to continue to address further or more deeply.[11]

After the case-control studies, the ABC epidemiologists sought to confirm some of their findings using a prospective design, and so planned the MEC cohort study in the 1990s. This "prospective cohort" included people with no known diseases who were to serve as subjects in multiethnic epidemiological studies to allow the study of the development of future illness and mortality. In total, about 215,000 participants were recruited as participants in the cohort.

> And we eventually got to the point where we felt we needed to have a cohort study to follow a large group of people over time to get better data from that perspective to associate different factors with cancer risk. That was around 1990. So [Director of ABC] did a tremendous amount of work, and then

11. Epidemiologist A, interview by author, Hawai'i, 2013.

> eventually was able to get funding for the initial multiethnic cohort study grant. So then the project actually began in 1993 . . . But we had already been involved with a cohort study with the Japanese ethnic group that began back in 1971.[12]

Because their early studies were interested in the relationship between diet/nutrition and disease, the ABC epidemiologists included ethnic groups that had different diets: Japanese-, Chinese-, Korean-, Filipino-, Native Hawaiian-, Samoan-, and Caucasian-Americans from Hawaiʻi.[13] As an MEC epidemiologist noted, "the early case-control studies included the five main ethnic groups in Hawaiʻi: Japanese Americans, whites, Native Hawaiians, Chinese, and Filipinos."[14] Hawaiʻi, and not the rest of the US, was their focus.

In summary, the MEC cohort took advantage of different ethnic groups with different diets co-located in Hawaiʻi. ABC researchers were adamant that they were interested in research on environmental variables like nutrition. Speaking in the vernacular of GxE studies (which attempt to disentangle gene-environment interactions), a lead scientist on the MEC study stated, "We can intervene in E, we cannot intervene in G, or at least not for a while."[15] He had initiated the MEC with an interest in diet and nutrition. When asked about changes in diet versus homogeneity in diet, he said:

> It's really the nutrients in the food that we were interested in, not the foods themselves. So it's the amount of fat, types of fat, etc. Japanese [most of the "Japanese" participants in the MEC study were second-generation Japanese-Americans] still tended to eat Japanese dishes. Whites eat sushi, but not as much Japanese food as Japanese families.[16] There have been increases in rates of cancers among Chinese in Southern China by 10 percent. This kind of sudden increase cannot be due to genetic changes, because genetics don't change that quickly. In Japan, rates of stomach cancer are now equivalent to rates in Hawaiʻi Japanese. Their diet has changed, increase in fat intake, meat intake.[17]

12. Ibid.

13. L. N. Kolonel et al., "A Multiethnic Cohort in Hawaii and Los Angeles: Baseline Characteristics," *American Journal of Epidemiology* 151 (2000): 346–57.

14. Epidemiologist A, interview by author, Hawaiʻi, 2013.

15. Ibid.

16. The diets of many of the MEC Japanese American research subjects tended to include dishes from the time of their parents' immigration in the early nineteenth century, though there had been changes in diet in Japan since then.

17. Epidemiologist A, interview by author, Hawaiʻi, 2013.

Everything Comes up Genomic: Transforming Environmental Risks to Genomic Risks

How did studies linking diet, nutrition, and cancer transform into studies linking race and ethnicity, SNPs, and cancer?

> The MEC wasn't set up to do genetic studies actually, because when we set it up, this whole era, this whole genomic era, hadn't really happened. It was coming, but it hadn't quite happened. So we set it up because we had a long-standing interest in nutrition and cancer, primarily, and had set it up to do, primarily, studies of diet, nutrition, and cancer.[18]

In the mid- to late-1990s, geneticists interested in using the outcomes of the Human Genome Project (HGP) began to build infrastructures for eventually conducting GWASs or statistical association studies in search of genomic risk factors for CCDs like cancer. In the section on reference populations, we describe how GWASs led to the construction of genomic categories onto which some writers placed ethnic and racial labels. The coalescence of several actors in this larger context led to the transformation of the prospective MEC database for the study of nutritional risk factors into a DNA database for use in these GWASs.

> We very quickly picked that up and decided that the MEC was the *perfect resource for doing studies of genetic susceptibility and cancer*. So, I guess about a couple of years into setting up the cohort, just before we had completely established the full cohort, we started to collect bloods from selected cancer sites, from people who'd developed breast, colorectal, and prostate cancer. We started collecting bloods from these *cases* and then we picked a cross-section of the cohort to serve as the *control* group, and that's how we started doing all of our genetic analyses, so it's all nested case-control studies. And we've continued to do that, but for a lot of studies for a lot of cancer sites. So, we were doing these nested case-control studies . . .[19]

> In the mid to late 1990s, we started to collect blood on recently diagnosed breast, prostate, colorectal cancer cases in the cohort and a group of non-cancer controls with the goal of doing genetic studies. At that time, those were candidate gene[20] or candidate pathway studies, not GWAS.

18. Epidemiologist B, interview by author, Hawai'i, 2010.

19. Ibid. (emphasis added).

20. Mid-twentieth century, human geneticists began developing methods for exploring the transmission of inherited conditions by examining chromosomes and genes. When an

> About five years later, we were able to get funded to collect blood and urine [from] all survivors who wanted to give a biospecimen through separate grants in Hawai'i [ABC] and in Los Angeles [XYZ]. Since those samples were collected before diagnosis, we were able to investigate association with biomarkers (e.g., hormones, diet, etc.), in addition to continuing to do genetic studies, which were then GWAS. After the [GWAS] biorepository was assembled, we continued to contact new cases who were not already in the biorepository to get a saliva sample by mail in order to increase our sample sizes for GWAS of the main cancer sites.[21]

The MEC then was established to confirm previous studies that linked nutrition with cancer. But with the advent of the HGP and GWASs, the MEC study biorepository was transformed into a database for seeking genomic or SNP factors involved in risk for disease. Statistical geneticists called this kind of search "unbiased" or "hypothesis-free" in that it sought SNP signals that might indicate genomic risk factors, independent of any a priori knowledge or hypothesis about where these SNPs might be in the genome, and without using any previous studies to inform or narrow down the search. Also referred to as "discovery-driven" studies, as opposed to "hypothesis-driven," such study designs have generated much debate about whether any research is truly hypothesis-free and uninformed by prior theory and data.

Over the last twenty years, 70,000 MEC study subjects contributed biospecimens (blood and urine samples), health measures, and information on diet and exercise practices. The GWAS included self-reported race for the participant and each of his/her parents, based on the race/ethnicity questions asked in baseline questionnaires.

Imposing Continental Diversity onto Local Diversity

In the process of transforming use of the MEC study biorepository from research on environmental risks for disease to research on genomic risks for

otherwise rare disease appeared across several generations in a family tree, researchers studied these families to identify markers and alleles of genes that were disproportionately present in affected individuals but not their unaffected relatives. Through this "candidate gene" approach, researchers zeroed in on the loci that seemed most promising as explanations for the familial inheritance of the disease under study. This approach identified many of the genes involved in Mendelian diseases.

21. Epidemiologist C, interview by author, Hawai'i, 2019. XYZ is a cancer research institute in Los Angeles.

disease, US OMB (Office of Management and Budget) racial categories came to dominate over the diversities local to Hawai'i.

> The concept, originally by myself and my colleague here, was to set up a multiethnic cohort in Hawai'i. But [an epidemiologist then head of the Epidemiological Section of XYZ] was visiting on sabbatical at about the same time, and the whole concept came up of cohorts, and he'd been thinking about cohorts because he'd just been in Japan and the Japanese were talking about a cohort. So we sort of were bouncing this around, and then we suggested—well, if we could include L.A., we'd be more *diverse*. Because we could add African Americans and a big Hispanic component which we don't have here [in Hawai'i].[22]

Beginning in 1993, the ABC epidemiologists joined forces with the XYZ in Los Angeles to conduct GWASs, and they planned to "include whites, African Americas, Latinos, Japanese Americans, Native Hawaiians, Koreans, and Chinese to be recruited in Hawai'i and California. The National Cancer Institute (NCI) reviewers asked that the number of groups be limited to the first four groups because of what they saw as too small a sample size for these other groups."[23] The NCI also told the MEC organizers that there were not enough Native Hawaiians, but the MEC director felt that they could not do a study in Hawai'i and not include Hawaiians, so they successfully sought another grant to fund their inclusion.[24] "[A cancer epidemiologist associated with the MEC] got a DOD [Department of Defense] breast cancer grant and added Native Hawaiians to the cohort."[25] In 1995–97, they collected blood for their candidate gene or candidate pathway studies.

> In the end, four ethnic groups (African-Americans, Japanese-Americans, Latinos, and Caucasians) were selected for the study based on the size of the respective populations in the two areas and on the striking differences among them in the reported rates for several common cancers. Subsequently, a smaller ethnic group, Native Hawaiians, was added.[26]

With the input of an epidemiologist from the XYZ in Los Angeles and at the behest of the NCI, "Latinos," "African Americans," and "Caucasians" from Los Angeles were added to create the five racial/ethnic groups that more closely

22. Epidemiologist B, interview by author, Hawai'i, 2010 (emphasis added).
23. Epidemiologist C, interview by author, Hawai'i, 2019.
24. Epidemiologist B, interview by author, Hawai'i, 2010.
25. Epidemiologist C, interview by author, Hawai'i, 2019.
26. Kolonel et al., "A Multiethnic Cohort" (ref. 13), 346–47.

emulated OMB and US Census race categories: "Asians" qua "Japanese Americans," and Native Hawaiians qua "Pacific Islanders," "Latinos," "African Americans," and "Caucasians." (The Caucasian category included participants from Hawaiʻi and Los Angeles.) This is the moment when US racial and ethnic categories, as mandated by the OMB and taken up by the NIH, transformed a project specific to Hawaiʻi and its different ethnic groups into a distinctly continent-based view of human diversity that mirrored the racialized lenses of the US census.

The notion that Hawaiʻi itself did not provide enough diversity, and that including African Americans and "Hispanic Americans"[27] would produce a "more diverse" cohort is based on a US continental perspective of what constitutes diversity. By including African American and Hispanic research subjects from Los Angeles, the MEC database was viewed as more "global" in its aim and reach, and its perceived relevance to human genomic studies elsewhere. But how is "global" represented? What does it mean for a localized study of diet and cancer to shift to a study of genomic markers and common diseases on a global level?

LINKING RACE AND ETHNICITY TO GENOMIC RISKS FOR DISEASE

US clinical studies in the mid- to late-twentieth century had typically included only "whites" (and primarily men), presumed to be the "representative" or ideal human biological type that could stand in for all human bodies. Studies that included individuals from US racial minority groups were rare, partly because these communities had long been subjected to medical exploitation.[28] In 1993, the NIH Revitalization Act, spurred by Director Bernadette Healey, prompted a slow shift to incorporating gender and racially diverse communities in clinical trials sponsored by federal taxpayer dollars.[29] The underlying rationale was twofold: A white-male paradigm of diagnosis and treatment had

27. For discussion about debates about this category in US Census, see Kenneth Prewitt, *What Is Your Race?: The Census and Our Flawed Efforts to Classify American* (Princeton, NJ: Princeton University Press, 2013).

28. E. Hammonds and S. Reverby, "Toward a Historically Informed Analysis of Racial Health Disparities Since 1619," *American Journal of Public Health* 109, no. 10 (2019): 1348–49.

29. Troy Duster, *Backdoor to Eugenics* (New York: Routledge, 1993); Steven Epstein, *Inclusion: The Politics of Difference in Medical Research* (Chicago: University of Chicago Press, 2007).

proven ineffective at diagnosing and treating the ailments of diverse humans (women in particular); dismantling the gender and racial homogeneity of clinical trial subjects would generate data and findings that were more comprehensive, representative of and applicable across the diversity of human biology, by taking account of biological difference. In addition, community health workers and activists,[30] and minority scientists,[31] advocated inclusion as a way to ensure that the health benefits of the HGP and what has been called the genetic revolution would benefit all Americans across the spectrum of race and gender.

Though this shift to diversity did not characterize early GWASs, which were still predominantly conducted on whites, nevertheless the large-scale genetic research projects that launched in the early 2000s, like the HapMap Project, actively sought to characterize genetic differences across more diverse sets of populations.[32] However, this was curiously counterbalanced by an assumption embedded in these early GWAS infrastructures: that CCDs would be underwritten by common disease-associated variants that would be found across all racial and ethnic groups (albeit at different frequencies), as conveyed in the mantra that guided the HapMap Project, "common disease, common variants."[33] The inclusion of genetic samples from diverse peoples aimed in part to confirm this assumption of commonality, though some were already arguing that diversifying genomics was necessary because significant differences might still be uncovered.[34]

30. Epstein, *Inclusion* (ref. 29); A. Nelson, *The Social Life of DNA: Race, Reparations, and Reconciliation After the Genome* (Boston: Beacon Press, 2016).

31. Duana Fullwiley, "The Biologistical Construction of Race: 'Admixture' Technology and the New Genetic Medicine," *Social Studies of Science* 38, no. 5 (2008): 695–735; R. Bliss, *Race Decoded: The Genomic Fight for Social Justice* (Palo Alto, CA: Stanford University Press, 2013).

32. This was partly because GWASs often used samples that had been collected years before for other medical research studies. See Joan H. Fujimura and Ramya Rajagopalan, "Different Differences: The Use of 'Genetic Ancestry' versus Race in Biomedical Human Genetic Research," *Social Studies of Science* 41, no. 1 (2011): 5–30.

33. M. W. Feldman and R. C. Lewontin. "Race, Ancestry, and Medicine," in *Revisiting Race in a Genomic Age*, ed. B. A. Koenig, S. S.-J. Lee, and S. S. Richardson (New Brunswick, NJ: Rutgers University Press, 2008), 89–101.

34. In contrast to this early predominating assumption, later thinking increasingly theorized the existence of rare variants, found at low levels (<5%) in most populations, but perhaps at higher, experimentally discoverable frequencies in a few populations, leading to the idea that different populations may well have different SNPs for a CCD.

CONSTRUCTING "REFERENCE" POPULATIONS

As previously noted, GWASs are statistical searches for genomic markers called SNPs for potential disease association. GWASs have served as the foundation for a global "bandwagon" phenomenon[35] shaping the landscape of twenty-first century human genetic research. They first aimed to find genomic risk markers for CCDs, and now have expanded to include the search for hypothesized genomic markers for human behaviors.

GWAS organizers hypothesized that there were points on the human genome where there are variations between individuals. These are called single nucleotide polymorphisms (SNPs, pronounced "snip") or simply "variants." A SNP is a single nucleotide in a genome sequence that differs between individuals. Each person has about three billion DNA base pairs in their genome, and a small percent of those DNA base pairs exhibit SNP differences. Most of these SNPs lie outside of the coding regions of genes, and most have no known function. GWASs originally aimed to search for genes for CCDs, the primary killers in nations where improvements in health and sanitation over the last hundred years, as well as changes in nutrition, have reduced mortality, increased lifespan, and precipitated an increase in rates of cancer, cardiovascular disease, Type II diabetes, and other chronic, age-related diseases. CCDs are not rare Mendelian diseases that are usually caused by single genes, but instead are likely a result of multiple risk factors, environmental and possibly SNP-related. Researchers working in early human genomics in the 1990s theorized that such SNPs might eventually lead them to genes.

Who initiated GWASs and why? As the HGP was nearing completion, NIH officials and medical geneticists proposed in the 1990s to study points of difference between individuals within their genomes, to aid in the search for genetic contributions to human disease. GWASs are statistical association studies based on algorithms that search for associations between SNPs and the risk of CCDs such as Type II diabetes, heart disease, etc.[36]

35. J. H. Fujimura, *Crafting Science: A Socio-History the Quest for the Genetics of Cancer* (Cambridge, MA: Harvard University Press: 1996).

36. GWAS finds *associations* between SNPs and disease but cannot establish causality. Thus, geneticists have tried to emphasize that neither SNPs found through GWAS, nor estimates for increased risk of CCD due to the presence of any particular SNP or set of SNPs, should be used for individual-level clinical prediction. Despite this, some private companies have been selling tests through the consumer marketplace that claim to tell individuals their risks for these diseases based on the presence of GWAS-identified SNPs.

In constructing statistical association technologies, some GWAS researchers worried about how to disentangle SNPs associated with disease risk from confounding SNPs that had nothing to do with disease. How could they identify and eliminate the effect of confounding SNPs from their analysis? Human population geneticists began to urge attention to "human population variations" as likely confounding culprits that could lead to spurious associations. In organizing their search for disease markers, human population geneticists argued that medical geneticists needed to "control for confounding SNPs" to "account for population substructure." On the basis of this advice, the medical geneticists decided to first construct "population differences" so that they could be accounted for in the statistical analysis.

In addition to ideas about racial and ethnic categories as having differences in inherent or genetic risk for disease, GWASs also introduced ideas about racial and ethnic categories as genomically differentiated categories in themselves. Leveraging the massive datasets generated by the HGP, genome researchers built durable, NHGRI-funded data infrastructures to facilitate statistical GWASs. The first GWASs were conducted using blood samples collected in Europe; NHGRI (and later NIH) director Francis Collins and Broad Institute founder Eric Lander established collaborations with Finnish investigators to enroll Finns in their initial GWASs under the assumption that Finns were a genetically homogenous population, thus tying ideas of national identity to ideas of "genetically isolated" populations.

Some of the GWAS researchers we studied did not see race as a relevant concept in their data analysis, asserting that race categories were an inaccurate, imprecise, and unscientific way of organizing and categorizing their data.[37] As noted earlier in this paper, based on human population genetic theories and historically varying patterns of exogamy and endogamy, GWAS organizers assumed that there were SNP differences between groups of people from different geographic regions of the world, and that these differences could be potential sources of spurious correlations that might arise in their statistical comparisons between research subjects with and without a given disease. They decided to "control" for these spurious correlations by "accounting for population substructure." Despite their insistence that race was not a salient set of categories for their research, it was at this point of

37. Fujimura and Rajagopalan, "Different Differences" (ref. 32).

controlling for population substructure that race was made durable in GWASs.[38]

To operationalize their "accounting for population substructure,"[39] medical geneticists employed theories from human population geneticists about early human movements. Population geneticists hypothesize that Africa was the home of the first humans, and that the ancestors of contemporary peoples on different continents left Africa in multiple waves beginning sometime between 200,000 and 110,000 years ago, migrating and settling across the globe.[40] There are few sources of data from that time period, in the form of bones and archaeological remains, but they theorize that different groups of people moved to what is currently Europe, Asia, across the ice bridge between what are now called Russia and Alaska, and down through the Americas.

Based on these ideas about early human migrations, medical geneticists constructed research infrastructures to enable them to collect, store, prepare, and analyze DNA for use in GWASs. One infrastructure was the HapMap Project. Medical and population geneticists proposed in the 1990s to extend the single composite genome being generated by the HGP, by identifying SNPs, points of DNA sequence difference between individuals, as markers for identifying genetic sources of disease risk. The HapMap Project organizers decided initially to collect DNA from groups in Ibadan (Nigeria), Tokyo (Japan), Beijing (China), and Utah. They delineated these samples by a combination of ethnic, geographical, and national identifiers; for example, they collected and genotyped DNA from 90 members of the Yoruba community in Ibadan, Nigeria, 45 Japanese in Tokyo, 45 Han Chinese in Beijing, and 90 Utah residents with northern and western European ancestry.[41] Though the

38. Geographic differences were transformed into "population" differences, as opposed to gradual, or clinal, variations across space.

39. The language of difference used by human geneticists is changing. In addition to "population substructure," they now also write that different populations have different "genetic architecture[s]"; G. L. Wojcik, M. Graff, K. K. Nishimura, et al. "Genetic analysis of diverse populations improves discovery for complex traits," *Nature* 570 (2019): 514–18, https://doi.org/10.1038/s41586-019-1310-4.

40. Some archaeologists have used 250,000 as an upper bound estimate, because of the technical uncertainties in dating bones and extracted DNA, and as new bone discoveries lead to revised dates of origin. The most current data based on found bones gives a 130–110K timeframe as the earliest migration estimate.

41. HapMap 3, a subsequent and final collection for the HapMap Project, added more samples from additional groups, but these groups were also labeled with a mix of ethnic and geographic identifiers that lent themselves to inferences of racial or continental origins, e.g., "Gujarati Indians in Houston, Texas" (cf. J. Hamilton, "Revitalizing difference in the HapMap:

HapMap Project groups were meant to be denoted by their specific ethnic, city, and national labels, they were later operationalized (by researchers who used HapMap data and downstream readers of GWASs) as representative of continental genetic variation (African, Asian, European).

Why did the HapMap Project planners choose to sample DNA from Yoruba in Ibadan, Japanese in Tokyo, Han Chinese in Beijing, and Utah residents with northern and western European ancestry? Did they not anticipate that these would be read as "Africans," "Asians," and "Europeans"—terms they strongly discouraged? Ethicists and social science researchers had warned that this configuration of collection sites would be viewed as representing racial categories and would be conscripted in attempts to use genetics to represent US racialized groups. They raised concerns that the HapMap Project could lead to ideas that racialized groups were genetically different from each other, and to ideas similar to those of nineteenth-century eugenicists about inherent differences between "races" in intelligence, behavior, and more.

In answer, geneticists argued that this was only their first round of reference population DNA collections, and that they would also collect DNA from many other groups in subsequent collections. (But these groups are also labeled with ethnic or racial terms, e.g., Gujarati Indians from Houston, Texas.)[42] Medical geneticists also responded that their initial choices of which groups to delineate for sampling would represent the most broadly diverse range of samples possible, within their pragmatic limitations of time, resources, and accessibility.

Thus, the HapMap Project designated groups that, because they were viewed as representing the genetics of particular continents (Asia, Africa, and Europe), became operationalized as continental reference populations and viewed by the biomedical community as representative of continental origins (e.g., the 45 "Japanese from Tokyo" samples and 45 "Chinese from Beijing" samples were seen as representative of genetic diversity in east Asia). The HapMap Project groups were initially denoted in HapMap by geographical (and for Yoruba, ethnic) labels, but via GWAS uses of the HapMap data, these labels were translated to continental populations that closely mirrored US census race categories. Importantly, when it came to controlling for substructure in GWASs, the researchers used the HapMap Project collection groups as

Race and contemporary human genetic variation research," *Journal of Law, Medicine & Ethics* 36, no. 3 (2008): 471–77.

42. Hamilton, "Revitalizing Difference" (ref. 41).

their points of reference, roughly corresponding to continents, to adjudicate whether self-identified study samples "really were" of the racial, geographic, or ethnic background their donors claimed. They did this by plotting and comparing the genomic variation found in study samples versus the HapMap samples, using statistical algorithms and software. Researchers viewed the samples whose genomic variants clustered more proximate to each other on these plots as more genetically similar and, importantly, inferred that they were more closely related ancestrally. In some cases, the genetic variation of an individual sample mapped far away from the rest of the samples whose donors had claimed a similar racial, ethnic, or geographic identity, and so, these mapping techniques were used to identify "outliers"—individuals who identified as belonging to one group, but their DNA clustered closer to another group. Outliers were removed from further analysis.

This stepwise translation, from HapMap sample groups to GWAS reference groups, through statistical software and analysis, created and mobilized "reference populations" within GWAS. During this process, researchers identified the SNPs whose variation seemed to contribute most to the separations among the reference populations on the cluster plots. These SNPs were often interpreted as group-specific or group-diagnostic SNPs (some of which were also known as ancestry-informative markers AIMs).[43] AIMs were seen as being able to "adequately control for the bias" of racial difference that marks "common race/ethnicity categories."[44] Thus, SNPs were directly conscripted into efforts to cast difference as racial, and racial difference as a confounding source of error in the GWAS.

Whether or not they believed that races are or are not genetic groups, the data infrastructures researchers built were organized by categories commonly associated with continents, in turn commonly associated with popular American notions of racial difference. Their clustering program has since been expanded, and many newer algorithms and software claim to perform the clustering even more precisely. Human population geneticists are publishing the plots made by these programs, which plots are often held by some

43. Fullwiley, "Biologistical Construction of Race" (ref. 31); Ramya M. Rajagopalan and Joan H. Fujimura, "Making History via DNA, Making DNA from History: Deconstructing the Race Disease Connection in Admixture Mapping," in *Genetics and the Unsettled Past: The Collision between DNA, Race and History*, ed. K. Wailoo, C. Lee, and A. Nelson (New Brunswick, NJ: Rutgers University Press, 2012), 143–63.

44. H. Wang et al., "Self-reported Ethnicity, Genetic Structure and the Impact of Population Stratification in a Multiethnic Study," *Human Genetics* 128, no. 2 (2010): 165–77.

geneticists as evidence confirming views that races map onto genomic categories of difference.

Human population geneticists and their programs "constructed" the populations. Geneticists do not have access to, and cannot work with, the DNA of people from 130,000 years ago. Instead, they use the DNA of contemporary peoples, as they exist today, to stand in for the DNA of people and groups at the time of migration out of Africa. Yet, through this time, humans have traveled, interbred, and shared DNA across populations, cultures, communities, nations, and continents. Ideas of national difference, of national origins, are built on ideas of purity within geographic areas. But ideas of purity are only ideas. There is virtually no data to substantiate them, because those human bodies no longer exist, except a few samples of DNA from a few bodies that have been preserved due to unusual conditions. Our goal here is not to undermine scientific efforts to understand human history, but to analyze how they are conducted, through which practices, using which tools and infrastructures, to understand how the outcomes are produced. We argue that the use of DNA of contemporary peoples to represent human migration patterns from over 100,000 years ago is built on extant sociocultural ideas of racial, ethnic, and national differences, projected onto the past.

To summarize, there were *contradictory* views of "race" operating in early twenty-first century genomic research on CCD. Some geneticists and epidemiologists believed that different groups displayed different genetic risks for CCD, that there were different disease variants for different "races," whereas others predicted that the differences would be insignificant compared to the common variants underlying these diseases. Some geneticists believed that individuals classified into racialized groups were "close enough" to each other and sufficiently different from individuals classified into other groups to serve as proxies for population categories.[45] But most of the medical geneticists we interviewed believed that US racial categories are not equivalent to genetic categories, and that using race categories would be incorrect science.[46] Nevertheless, continentally defined reference populations are integral parts of GWAS.

45. E. Burchard et al., "The Importance of Race and Ethnic Background in Biomedical Research and Clinical Practice," *New England Journal of Medicine* 348 (2003): 1170–75; D. Reich, *Who We Are and How We Got Here: Ancient DNA and the New Science of the Human Past* (New York: Pantheon, 2018).

46. Stephen Jay Gould, "The Geometer of Race," *Discover* 15, no. 11 (1994): 64–69; Frank Livingstone, "On the Non-Existence of Human Races," *Current Anthropology* 3 (1962): 279–81; R.

DISCUSSION AND CONCLUSIONS

Constructing "Population Differences" using HapMap Data and MEC Data

Race came to big data GWASs through the work of statistical geneticists who aimed to avoid spurious correlations in their statistical analysis by relying on human population geneticists' ideas about genomic distances among racial and ethnic groups.[47] By using particular communities on three continents to account for presumed human population differences, geneticists allowed racial and ethnic categories to be built into GWAS infrastructures and impact the construction of human population genomic categories. Thus, ideas about how to collect the initial HapMap samples show how race, nation, and ethnicity are fused with ideas about DNA differences.[48]

The initial HapMap collections have become entrenched in human population genetic algorithms and software in ways that have long-term ramifications on notions of race and ethnicity in the US and abroad. Human population geneticists associated with the HapMap Project and subsequent GWASs developed algorithms and software, identifying and characterizing SNPs collected from Africa, Asia, and Europe in ways that highlighted and racialized the differences between those DNA collections. Their algorithms characterized *patterns* of genomic variation in humans. The software compares SNP patterns in the genomes from the study subjects with SNP patterns in the HapMap reference populations and graphically depicts the similarities and differences its algorithms find. These algorithms decide the boundaries of populations. Because the SNP variants selected by the HapMap Project were collected from individuals characterized by continental uniqueness and national origin labels, labels of race and ethnicity quickly became attached to the constructed

C. Lewontin, "The Apportionment of Human Diversity," in *Evolutionary Biology*, ed. T. Dobzhansky, M. K. Hecht, and W. C. Steere (New York: Springer, 1972). As Gould (1994) and Livingstone (1972) had argued, human genetic variation is clinal and cannot be organized into groups of individuals along any characteristics, and as Lewontin (1972) demonstrated, within-racial-group genetic differences are greater than between-racial-group genetic differences. See J. H. Fujimura et al., "Clines without classes: How to Make Sense of Human Variation," *Sociological Theory* 32, no. 3 (2003): 208–27, for elaboration on this issue with respect to the use of human population genetic technologies.

47. See Duster, *Backdoor to Eugenics* (ref. 29) for other ways in which race came in through the back door.

48. S. S.-J. Lee, "Racializing Drug Design: Implications of Pharmacogenomics for Health Disparities," *American Journal of Public Health* 95, no. 12 (2005): 2133–38; Hamilton, "Revitalizing difference" (ref. 41).

HapMap reference populations and their SNP patterns, and mapped onto the genomes of the MEC study populations ("Native Hawaiian," "Japanese American," African American," "Latino," and "Caucasian").

Indeed, industry, academic, and government genome scientists developed high-throughput technologies (like SNP microarrays, aka chips) to assess genomic variation, based on collections from the HapMap Project. These SNP chips were later explicitly redesigned for use only with populations circumscribed in racial and ethnic terms. In other words, each population came to be seen as best assayed by a SNP chip built specifically with SNPs chosen to represent the frequencies of "their" haplotypes and genomic variants. But "their" hapolytpes and genomic variants were based on SNPs selected from already socially designated groups. In other words, socially designated groups were used to define "biological" groups.[49]

Our argument is that, no matter their intent, geneticists have *built* practices and materials that use social notions of race, nation, and ethnicity for distinguishing groups *into* each layer of their genomic research infrastructures. This *matters*. For example, *socio-material practices*[50] determine what one sees in the clusters produced by the statistical clustering software. These clusters are not made in nature. Scientific concepts of materiality in this case were determined by the researchers' decisions and assumptions about human genomic variation. These decisions and assumptions were key factors that determined the outcome.

The history of the MEC study demonstrates *how* non-white populations were developed for and used *first* in epidemiological studies of environmental factors involved in disease risk. The population samples and data were *later* used to search for genomic contributions, like SNPs, as risk factors for disease. This turn itself echoes larger shifts in studies of human disease that took place as genetics and genomics developed in the late twentieth century and began to seep into every corner of biomedicine, from consideration of outward ("social" or "environmental") factors involved in disease, to a focus on innate (genetic or biological) factors involved in disease. Some of the MEC organizers wanted eventually to be able to conduct "GxE" studies that could identify interactions between genetic susceptibility and lifestyle/environmental factors. Their hope

49. See Ramya M. Rajagopalan and Joan H. Fujimura, "Variations on a Chip: Technologies of Difference in Human Genetics Research," *Journal of the History of Biology* 51, no. 4 (2018): 841–73.

50. Joan H. Fujimura, "Sex Genes: A Critical Socio-material Approach to the Politics and Molecular Genetics of Sex Determination," *Signs* 32, no. 1 (2006): 49–82.

was that building the MEC biorepository would lead to such GxE studies.[51] However, there were several unintended consequences of the combined use of race- and nation-identified groups in the collection and analysis of *reference populations*, on the one hand, and the selection of individuals from racial-, ethnic-, and national-identified groups for *study populations*, on the other.

The first unintended consequence is that some geneticists (like David Reich),[52] science writers (like Nicholas Wade),[53] and members of the public (including leading white supremacists like David Duke and others) now interpret the population plots generated by human population genetics software developed for GWASs, as evidence that racial groups are genomically different categories.[54] As we have discussed, they do not take into account the fact that social ideas of race played significant roles in the *construction of human genomic categories*.

But could it have been otherwise? There were other options available to researchers in place of using ethnic identifiers.[55] For example, instead of referring to collections by continental labels, designers of software for genomics analysis could have opted to use labels for the different collections that had no affiliation with nation, tribe, or continent. Both producers and users could instead have referred to HapMap group collections in terms of their DNA similarity. For genetic studies of disease, it is only DNA similarity that matters, not any inferences about why that similarity exists or where it came from. After controlling for such similarities without appeals to racialized labels, researchers could have continued to look for SNP differences that

51. "The search for the causes of cancer and means of cancer prevention has entered a new era as recent developments allow correlation of environmental and behavioural exposures, genetic variation and patient outcomes. The Multiethnic Cohort Study was designed to take advantage of these advances to prospectively explore the roles of lifestyle and genetic susceptibility in the occurrence of cancer. The ethnic diversity of the cohort in this study provides a wide range of dietary exposures and genetic variation, thereby providing a unique dimension to this research." L. N. Kolonel, D. Altshuler, and B. E. Henderson, "The Multiethnic Cohort Study: Exploring Genes, Lifestyle and Cancer Risk, *Nature Reviews Cancer* 4 (Jul 2004): 1–9.

52. Reich, *Who We Are* (ref. 45).

53. Nicholas Wade, *A Troublesome Inheritance: Genes, Race and Human History* (New York: Penguin Press 2014).

54. For more on this, see J. Kahn et al., "How not to talk about race and genetics," *Buzzfeed* (30 Mar 2018).

55. Pragmatist Interactionist perspectives in sociology argue that social organizations are not set in stone, that there are many possible forms, that any form could have been otherwise. Everett Hughes gives Robert Musil credit for this; see Everett Hughes, *The Sociological Eye* (Chicago: Aldine Atherton, 1971). This perspective also applies to the organization of scientific work.

might figure in disease. For example, an early HapMap organizational option had been to use biologist Allan Wilson's plan to collect DNA from one person per square mile, from a virtual grid thrown around the world. This would have allowed DNA collection to happen without group labels. The work that would have enabled this plan was enormous, so the planners rejected it. But had they put in the front-end effort needed for this plan, they potentially could have eliminated the back-end work of dealing with the racializing consequences of the HapMap.[56]

Haraway posited three themes, each of which has dominated different time periods: race for the period 1900–1930s, population for the period 1940–1970s, and genome for period 1975–1990s.[57] She posed these concepts as general organizing themes in biology and society, and did not assume that the dominant theme governs all ideas in each era. *Our argument* is that not only are the three still equally dominant in twenty-first century genomics, they are also co-constructed. The second major consequence is that nineteenth-century ideas of race category differences have been used to construct both populations and genomic categories, thereby making it appear as if racialized groups constitute different genomic categories. We have presented examples of scientists' thinking of a human as an admixture of race categories, of GWAS study populations and HapMap reference populations using ethnic-, national-, and city-labeled groups of humans as sources of genomic biomaterials for constructing categorical differences. By examining the practices of scientific work instead of only statements made by scientists, we show how "race" has become imbricated in human population genetics and then into genomic biomedicine. We have shown how locally based epidemiological study populations became conflated with HapMap reference populations through a series of choices that tell us how scientists have made complex judgments about significance, scope, proxies, diversity, and how to use the possibilities and limits of algorithmic software tools for processing genomic data.

Constructing the Statistical Genomic Research Subject

Social scientists have framed the concept of "whiteness" as the *generic, unmarked category* in US society, where the notion of the unmarked category

56. Efficiency in scientific work often does not take into account its potential social consequences; see Joan H. Fujimura, "Constructing 'Do-able' Problems in Cancer Research: Articulating Alignment," *Social Studies of Science* 17, no. 2 (1987): 257–93).

57. Haraway, *Modest_Witness* (ref. 1).

refers to, in this case, the norm.[58] Whiteness has constantly been under construction, being remade in different eras under different conditions, especially in the context of colonization and imperialism.[59] But, in the US and other colonized territories, that construction has been positioned asymmetrically in relation to other racialized groups and cultures. Whiteness has been remade as a location of structural advantage, and as a site in which certain cultural practices and social identities are pegged as "normal" and, in the case of the US, have come to define structurally what constitutes the "national" (the "American") identity.

Discovery-driven research strategies such as GWAS had focused, in their early studies, on white research subjects because geneticists could collect enough biological samples and health information from "whites" in order to conduct big data, large-scale studies. But another reason is that NIH grant reviewers sometimes encouraged investigators to focus only on "whites" on the assumption that they would be from the same "population." They equated "race" with "population," and theorized that whites constituted one race, and were genomically "similar enough" to each other to avoid confounding. This assumption about within-race similarity is at odds with Lewontin's demonstration[60] that genetic differences within racial groups are greater than genetic differences between those groups. That assumption is further challenged by our research, which shows that the production of population differences is based on decisions and choices made during the scientific work processes—decisions made using social assumptions about racial differences.

After the initial big data, discovery-driven GWASs, including DNA from people of non-white populations became increasingly valorized. The MEC was

58. R. Frankenberg, *White Women, Race Matters: The Social Construction of Whiteness* (Minneapolis: University of Minnesota Press, 1993); R. Frankenberg "The Mirage of an Unmarked Whiteness," in *The Making and Unmaking of Whiteness*, ed. Birgit Rasmussen, Eric Klinenberg, Irene Nexica, and Matt Wray (Durham, NC: Duke University Press, 2001), 72–79; George Lipsitz, *The Possessive Investment in Whiteness: How White People Profit from Identity Politics* (Philadephia: Temple University Press, 2006).

59. M. F. Jacobson, *Whiteness of a Different Color: European Immigrants and the Alchemy of Race* (Cambridge, MA: Harvard University Press, 1999); David R. Roediger, *Working Toward Whiteness: How America's Immigrants Became White; The Strange Journey from Ellis Island to the Suburbs* (New York: Basic Books, 2005); L. A. Stoler, *Race and the Education of Desire* (Durham, NC: Duke University Press, 1995); Cheryl Harris, "Whiteness as property," *Harvard Law Review* 106, no. 8 (1993): 1707–91.

60. Lewontin, "Apportionment of Human Diversity" (ref. 46); Fujimura et al., "Clines without Classes" (ref. 46).

among the first large-scale genomic cohorts to include "multiethnic" participants whose data (on physical activity, diet, medical history, and ethnicity) and biomaterials (blood, DNA, and urine samples) could be channeled into prospective studies on genetics and disease risk. The MEC project benefited from this interest in diversity in GWASs, because the MEC was a rare repository of biomaterials from non-white participants. The MEC DNA was used in other GWASs in which MEC researchers collaborated. But the consequences of incorporating this notion of diversity into GWASs require careful consideration. Their "diversity" was designed both against the prevailing white norm and in terms of the NIH's commitment to using OMB categories for representing and including racial and ethnic diversity in *clinical* research.[61] To be clear, GWAS organizers and human population geneticists were interested in *genomic diversity in basic research* on risk for disease. Their interest in OMB categories was about the ability to capture genomic differences across the globe. The MEC researchers had planned to collect samples from Chinese-Americans, Filipino-Americans, Korean-Americans, and Samoan-Americans living in Hawai'i, but dropped those plans and instead adopted ethnic group categories that better fit the OMB categories because of their collaborations with GWAS organizers and human population geneticists. The final MEC study populations were framed as self-identified Japanese Americans, Native Hawaiians, Latinos, African Americans, and Caucasians/whites. In this case, the MEC was made more "diverse" according to human population geneticists' ideas of global geographic population diversity through the addition of African Americans and Latinos in Los Angeles to the ethnic groups present in Hawai'i, where there were fewer African Americans and Latinos.

After the diversion to genomic risk searches, the MEC study mirrored the HapMap 1 populations, and allowed researchers to interpret results of genomic studies based on MEC data to be broadly applicable to continental populations across the globe. That is, the DNA of these individuals in Hawai'i and Los Angeles was supposed to represent the DNA of groups across the globe. This out-of-place globalization of what was initially a very local cohort for studying nutrition and cancer, is in line with prevailing but overly simplifying

61. For earlier discussions of the politics surrounding the OMB categories and institutional mandates to include people of color in clinical studies, see Duster, *Backdoor to Eugenics* (ref. 29), and Epstein, *Inclusion* (ref. 29). For different perspectives on "adding diversity," see Wojcik et al., "Genetic analysis" (ref. 39), and S. Knerr et al., "Inclusion of Racial and Ethnic Minorities in Genetic Research: Advance the Spirit by Changing the Rules?," *The Journal of Law, Medicine and Ethics* 39, no. 3 (2011): 502–12.

understandings of DNA as immutable, universalizeable, and representative across time and place. "Our hope is that this knowledge will form a foundation for developing improved diagnostic, therapeutic and preventive interventions for broad use across the world's population."[62]

The turn to genomics in the work of the MEC study has had an impact on the way that other researchers in the broader field of human genomics think about the salience of race and ethnicity to and within genomics research. The use of the MEC biorepository has facilitated transitions in human genomics, around the idea that pre-existing GWASs and resources for human genomics were "too white" and not sufficiently representative of human biological (or genetic) diversity. Recently, the GWAS literature has exploded in the number of studies that claim to be ameliorating this gap, using terms like "multiethnic GWAS," "multi-ancestry GWAS," or "trans-ethnic GWAS." MEC research provided a model that has facilitated the emergence of and solidified these commitments to diversity in human genome research, which nevertheless put human genomics onto a path that increasingly ethnicizes and racializes study participants and groups.[63]

Since the HGP, and especially in the last ten years, there have been increasing calls to build DNA datasets that include people of color in order to align genomics with health disparities research. In the 1990s, discussions of race in the context of the human genome began to raise concerns flagged by social scientists and bioethicists.[64] Health disparities among different racial and ethnic groups and associated risk factors in the United States had been examined since well before the 1990s, indeed, since the early days of colonial America. Causes for these disparities had been located "in the body alone or in social conditions." [65] In the social sciences, at least in recent history, most of those disease risk factor studies had focused on socioeconomic and

62. Kolonel et al., "The Multiethnic Cohort Study" (ref. 51). L. N. Kolonel, D. Altshuler, and B. E. Henderson, "The Multiethnic Cohort Study: Exploring Genes, Lifestyle and Cancer Risk, *Nature Reviews Cancer* 4 (July 2004): 1–9.

63. More recently, other longitudinal population health repositories have followed a similar model to assay health differences and similarities across ethnic backgrounds, such as the Singapore Multiethnic Cohort study. See Kristin Hui Xian Tan, Linda Wei Lin Tan, Xueling Sim, et al., "Cohort Profile: The Singapore Multi-Ethnic Cohort (MEC) Study," *International Journal of Epidemiology* 47, no. 3 (2018): 699–699j.

64. Duster, *Backdoor to Eugenics* (ref. 29); P. Ossorio and T. Duster, "Race and Genetics: Controversies in Biomedical, Behavioral, and Forensic Sciences," *American Psychologist* 60, no. 1 (2005): 115–28.

65. Hammonds and Reverby, "Toward a Historically Informed Analysis" (ref. 28).

environmental risk factors.[66] But the movement toward genetic or genomic risks for disease created the potential to locate risk not only in individual bodies, but also in the racialized bodies of large groups of people.

The confluence of statistical epidemiological reference populations, like the HapMap 1 samples, with traditional epidemiological study populations, like the MEC groups, had *consequences*. The use of MEC populations in genomic disease risk studies brought theories and methods from human population genetics and statistical genetics to bear on studies of the bodies of MEC's ethnically classified individuals. These bodies were fashioned into statistical genetic entities, just as samples from them were transformed into resources for statistical genetic aims. Concepts of ethnic diets based on foods brought from home cultures became transformed into race categories qua human population genomic categories, as environmental causes of disease became diverted instead toward searches for genomic risk factors for disease.

This article has highlighted processes of what we call *genomic racialization*. As noted, Hawai'i had been populated by Hawaiians until the armed takeover by American businessmen. Thereafter, peoples from other parts of the globe have settled there, and Hawai'i now includes many diverse populations. However, it is not a microcosm of the US. Differences in Hawai'i do not mirror US continental diversities. African Americans and Latinos from Los Angeles were imported into the MEC in part to conform with the framework established by the US OMB categories. Just as US race categories were brought to Hawai'i through annexation,[67] the domination of US OMB race categories over a study cohort of local ethnic groups begun in Hawai'i is yet another colonization that has contributed to the racialization of the genome. Early studies of nutrition and cancers had a rationale for using ethnicity because different ethnic groups brought different diets to Hawaii. These

66. Even environmental risk factors like diet and nutrition and lack of exercise can be used to "blame the victim," when socioeconomic factors and systemic racism are the fundamental causes of disease. For alternative approaches to health inequality research, see N. Krieger, "Proximal, Distal, and the Politics of Causation: What's Level Got to Do With It?," *American Journal of Public Health* 98, no. 2 (2008): 221–30; Arline T. Geronimus et al., "Jedi public health: Co-creating an identity-safe culture to promote health equity," *Social Science and Medicine—Population Health* 2 (2016): 105–16. See also T. Zuberi and E. Bonilla-Silva (eds), *White Logic, White Methods* (Lanham, MD: Rowman & Littlefield Publishers, 2008), and T. Zuberi, *Thicker Than Blood: How Racial Statistics Lie* (Minneapolis: University of Minnesota Press, 2001) for a critique of the use of race classification and statistics in the social sciences, where the consequences become attached to racialized groups rather than to structural racism.

67. Domínguez, "Exporting U.S. Concepts of Race" (ref. 5).

earlier ethnic categories became co-opted in the interest of building the multiethnic cohort to study genomic risks for disease.

As environmental risks gave way to genomic risks, ethnicity gave way to race. Our concern is that the racialization of peoples of color in the US appears now to be developing into a racialization of the genome itself. Many medical geneticists and physicians, including people of color, have called for diversifying the genome, with the idea of bringing the "benefits" of genomics research to peoples of color. But there are unintended consequences of racially diversifying the genome. A consequence of the geneticization of race during the HGP is the *racialization* of the *genome*. Many have argued that GWAs have not thus far yielded much, especially given the amount of funding they received. But, for example, they have produced the finding of the 8q24 variants related to prostate cancer risk. An epidemiologist in our study, who said that "8q24 sits on the African portion of the genome," understood that this statement refers to the idea that certain variants in the genome "originated in Africa" and "traveled with" the groups migrating out of Africa to Europe, Asia, and the Americas. They understood that variants at 8q24 exist in all populations at different frequencies. Nevertheless, some researchers think of prostate cancer as an African American disease, even though the variants at 8q24 exist in people who self-identify as white, Asian, or Indigenous peoples in the Americas. Mobilizing racialized groups as racialized categories in genomic research to "diversify the genome" has impacts on how people read social categories into the genome.[68] That is, they may view the genome as divisible into racialized portions. This opens the door to "blame the victim" outcomes whereby diseases become associated with particular racialized groups used to study those diseases.

ACKNOWLEDGEMENTS

This article has benefitted from comments from Warwick Anderson, Susan Lindee, Loic le Marchand, and two anonymous referees. We are grateful to the scientists, researchers, and Hawaiian Studies scholars we interviewed during many years of this research. This research was supported by NSF Grant #1755003.

68. D. Nelkin and S. Lindee, *The DNA mystique: The gene as a cultural icon* (New York: Freeman, 1995).