MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190369

Racial Representation Disparity of Population-Level Genomic Sequencing Efforts

Isaac E. Kim, Jr., Indra Neil Sarkar

Center for Biomedical Informatics, Brown University, Providence, RI

Abstract

To develop personalized treatments for diseases, it is essential that they reflect the population of individuals that may be affected by a given disease. Amidst claims that there may be racial disparities in research populations, there have been no direct studies to explore this disparity in disease incidence and research projects that involve genomic sequencing. The precise relationship between underrepresentation of certain races in genomic sequencing studies and health outcomes relative to these races is unknown. Here, we examine the disparities in racial representation of national datasets pertaining to clinical data, mortality rates, and a major initiative involving genomic sequence analysis (The Cancer Genome Atlas [TCGA]). The results suggest that black Americans are underrepresented for most cancers in TCGA compared to clinical and mortality datasets, whereas Asian Americans are overrepresented. These findings accentuate the importance of targeted efforts to recruit representative patient populations into studies involving genomic sequencing.

Keywords:

Precision Medicine; Healthcare Disparities; Cohort Studies

Introduction

Genomic sequencing initiatives such as The Cancer Genome Atlas (TCGA) aim to catalogue cancer-associated genetic mutations towards the overall goal to diagnose, treat, and prevent cancer through better genetic understanding. While different genetic variants may result in similar symptoms, they could lead to diseases that require distinct, "personalized" treatments [1]. Genomic sequencing studies have led to major breakthroughs in the understanding of various types of cancer. In particular, research using TCGA has revealed that the genetic mutations responsible for breast cancer can be categorized into four major subtypes.

As precision medicine initiatives are embarked upon, such as the All of Us Research Program in the United States, it will be essential to be aware of the potential gaps in genetic knowledge. The All of Us Research Program is designed to treat patients based on individual differences in lifestyle, environment, and biology including factors such as race. Studies have shown that genetic makeup across races can impact treatment regimens as well as outcomes. For example, some patients with localized prostate cancer are prescribed active surveillance as opposed to immediate treatment. Other studies have shown, however, that Black American candidates for active surveillance had worse clinicopathological features on final surgical pathology than their white counterparts, suggesting that the criteria for active surveillance should be more rigorous for Black Americans [2].

The advancement of precision medicine requires genetic information on patients of all races [3]. Adequate racial representation in these studies will lead to more effective targeted therapies and at least address the issue of racial disparities in national health measures. Previous studies have shown a disparity between genomic databases and population demographics [4].

This study aimed to quantify the level of representation of racial minorities such as Black Americans and Asian Americans in the genomic sequencing study, TCGA, relative to clinically reported populations in three publicly available epidemiology databases: (1) Health Care Utilization Program [HCUP] from the Agency for Healthcare Research Quality; (2) mortality data from the Centers for Disease Control and Prevention [CDC]; and (3) the National Cancer Institute's Surveillance, Epidemiology, and End Results Program [SEER]). A major goal of this study is to provide data to compel researchers leading cohort-based studies to actively ensure appropriate balance of racial minorities.

This study sought to address the knowledge gap in the representation of racial minorities in genomic studies compared to incidence databases. To quantify and assess the proportional relationships between the racial groups examined in this study, we introduce the Sinuosity Index as a measure to quantify racial discordance.

Methods

Datasets

Four datasets were analyzed and compared: (1) the 2012 National Inpatient Sample (NIS) from the Healthcare Cost and Utilization Project (HCUP), (2) 2012 mortality data from the Centers for Disease Control and Prevention's (CDC) National Vital Statistics Report, (3) genomic sequencing data from TCGA, and (4) 1973-2014 incidence data from the National Cancer Institute's Surveillance, Epidemiology, and End Results Program (SEER). Based on the Clinical Classifications Software (CCS) disease codes that aggregate International Statistical Classification of Diseases and Related Health Problems, patient race information was extracted according to cancer groups from HCUP and SEER.

HCUP included the races White, Black, Hispanic, Asian or Pacific Islander, Native American, Other, and Invalid, while SEER catalogued a patient's race as White, Black, American Indian, as well as a range of ethnicities belonging to the Asian or Pacific Islander demographic, Other, or Unknown. The CDC's mortality data were analyzed from its corresponding website, while TCGA's genomic sequencing data was retrieved from the National Cancer Institute's Genomic Data Commons Data Portal. CDC divided race into White, Black, American Indian or Alaska Native, and Asian or Pacific Islander, whereas TCGA listed the races White, Black or African American, Asian, American Indian or Alaska Native,

Native Hawaiian or Other Pacific Islander, Other, and Not Reported. As a result of the different race variables and data formats across the four databases, the data extraction code required mapping tailored to each database. Tallies were graphed according to overall cancer and cancer, gender, and race for white, black, and Asian Americans using the Plotly.jl Julia package. This study included 16 cancers in total, specifically cancers of the head and neck (CCS code 11), esophagus (CCS code 12), stomach (CCS code 13), colon (CCS code 14), liver and intraheptic bile duct (CCS code 16), pancreas (CCS code 17), bronchus/lung (CCS code 19), skin melanomas (CCS code 22), breast (CCS code 24), uterus (CCS code 25), cervix (CCS code 26), ovary (CCS code 27), prostate (CCS code 29), bladder (CCS code 32), kidney and renal pelvis (CCS code 33), and brain and nervous system (CCS code 35). For certain cancers, data only existed in a single incidence dataset, SEER, leading to the omission of the following codes in the plotted graphs: Other non-epithelial cancer of skin (CCS code 23), Cancer of other female genital organs (CCS code 28), Cancer of other male genital organs (CCS code 31), and Cancer of other urinary organs (CCS Code 34). Analysis of Variance (ANOVA) calculations were done based on the total number of individuals for each race in each dataset using the Tukey honest significant difference test.

Sinuosity Index

Sinuosity index, which is a measure of steepness of a curve commonly used to assess the straightness of geographic features [5] was used as a measure for racial discordance across the examined databases. The relative difference was quantified as the slope between the relative lowest and highest occurrence of a given population group, with higher sinuosity indices suggesting greater variation across the databases.

For each given cancer and race in which at least three datasets were represented, percentage values were sorted in ascending order into an array. The highest value was then assigned an adjusted value of 1, while all other values were divided by the highest value. For the adjusted sinuosity index of each examined cancer and race, in traversing a sorted array of three values, the lowest value was given an x-value of 1 and set to the front index of a new 2D array, the second an x-value of 3 and set to the last, and the third an x-value of 2 and set to the middle index. In traversing a sorted array of four values, the lowest value was given an x-value of 1 and pushed to the front index of a new 2D array, the second an x-value of 4 and set to the last index, the third an x-value of 2 and to the second index, and the fourth an x-value of 3 and set to the third index. The euclidean distance between the first and last points in the sorted 2D array was calculated and designated as the B value. The sum of the euclidean distances between neighboring points (i.e., first and second, second and third, third and fourth, etc.) in the sorted 2D array was calculated and designated as the A value. The sinuosity index was calculated as A/B.

For the adjusted slope angle of each cancer and race, the lowest value was given an x-value of 1, the second lowest an x-value of 2, and so forth. Then, the euclidean distance was calculated between the points with the lowest and highest y-values and designated as the D value. Next, the difference between the points with the lowest and highest x-values was calculated and designated as the C value. The slope angle was

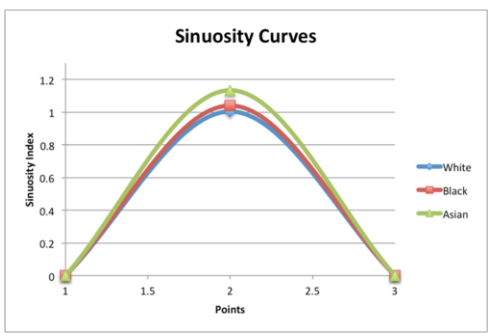

Figure 1 – Sinuosity Curves. Disproportion between sinuosity indices for White, Black, and Asian Americans can be seen as relative differences between the curves.

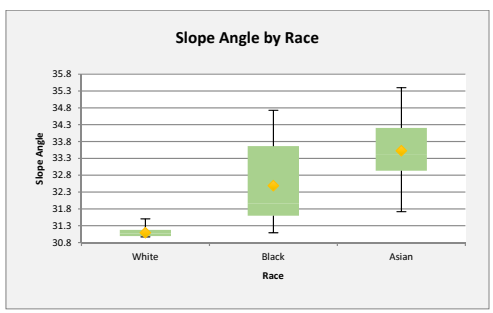

Figure 3 – Distribution of Slope Angle By Race. A distribution of slope angle is shown for each race examined across the datasets: White, Black, and Asian

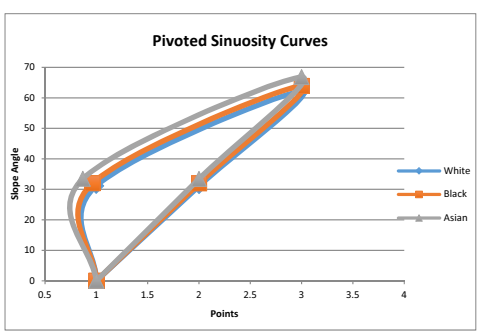

Figure 2 – Pivoted Sinuosity Curves. The relative slope differences between the White, Black, and Asian Americans shows disparity between genomic and epidemiological data sources.

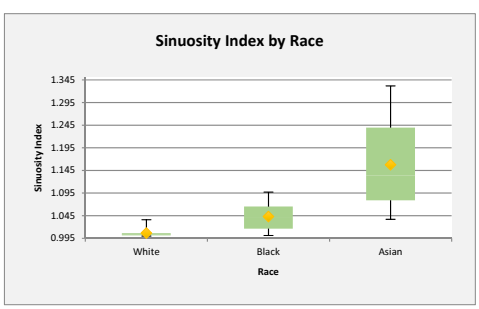

Figure 4 – Distribution of Sinuosity Index By Race. A distribution of sinuosity index is shown for each race examined across the datasets: White, Black, and Asian

calculated using cos(C/D).

The sinuosity curve was plotted for each racial grouping based on the median sinuosity index and then pivoted according to its respective median slope angle. The sinuosity curve was graphed by assigning two points with coordinates (1,0) and (3,0) and a third point with an x-value of 2 and y-value of the median sinuosity index. The sinuosity curves were pivoted so that the line connecting the points (1,0) and (3,0) corresponded to the median slope angle.

Results

For the cancers included in TCGA, Black Americans comprised less of the total sample population in TCGA as compared to HCUP, CDC, and SEER for 15 out of 19 cancers; White Americans and Asian Americans were underrepresented in 11 and 4 out of 19, respectively. Prostate cancer showed the most underrepresentation in TCGA across all examined races, and certain female-dominated cancers such as breast and uterus showed the highest representation of Black Americans in TCGA compared to HCUP, CDC, and SEER. An Analysis of Variance test using Tukey's honest significant difference test for the entire populations in each database revealed statistically significant differences (P=0.0000) between each race group for each dataset, except for Black Americans versus Asian Americans in SEER (P = 0.4001).

For the 16 cancers in which data in TCGA and at least two other datasets were listed, the mean sinuosity indices for White Americans, Black Americans, and Asian Americans respectively were 1.00642±00879 [1.00023-1.03608], 1.04298±02936 [1.00128-1.09715], and 1.15744±09114 [1.03703-1.33156]; the mean slopes were respectively 31.0961±0.1401 [30.9656-31.5034], 32.4931±1.22428

Table 1 – Sinuosity Indices and Slope Angles by Cancer and Race. Percent-adjusted sinuosity indices and percent-adjusted slope angles are shown for each cancer and race.

Cancer (CCS Code)	Race	Percent-Adjusted Sinuosity Index	Percent-Adjusted Slope Angle (Degrees)	Cancer (CCS Code)	Race	Percent-Adjusted Sinuosity Index	Percent-Adjusted Slope Angle (Degrees)
Head and Neck (11)	White	1.00438	31.0737	Breast (24)	White	1.00198	30.9822
Head and Neck (11)	Black	1.01063	31.3418	Breast (24)	Black	1.04249	31.6611
Head and Neck (11)	Asian	1.12017	33.1793	Breast (24)	Asian	1.03703	31.9721
Esophagus (12)	White	1.01167	31.2422	Uterus (25)	White	1.00066	30.9778
Esophagus (12)	Black	1.06297	33.6537	Uterus (25)	Black	1.03717	32.2554
Esophagus (12)	Asian	1.33156	35.3975	Uterus (25)	Asian	1.06960	32.7279
Stomach (13)	White	1.00507	31.1776	Cervix (26)	White	1.00128	30.9941
Stomach (13)	Black	1.07589	34.3908	Cervix (26)	Black	1.06337	31.9875
Stomach (13)	Asian	1.14658	33.5137	Cervix (26)	Asian	1.06368	32.9994
Colon (14)	White	1.00142	30.9756	Ovary (27)	White	1.00344	31.0783
Colon (14)	Black	1.03610	31.5895	Ovary (27)	Black	1.01553	31.7026
Colon (14)	Asian	1.08045	31.7173	Ovary (27)	Asian	1.15474	33.2923
Liver and Intraheptic Bile Duct (16)	White	1.03608	31.5034	Prostate (29)	White	1.01355	31.1755
Liver and Intraheptic Bile Duct (16)	Black	1.04369	33.2769	Prostate (29)	Black	1.08430	33.8490
Liver and Intraheptic Bile Duct (16)	Asian	1.24043	34.5932	Prostate (29)	Asian	1.23926	34.0319
Pancreas (17)	White	1.00770	31.1781	Bladder (32)	White	1.00247	31.0048
Pancreas (17)	Black	1.07147	33.6909	Bladder (32)	Black	1.01656	31.1584
Pancreas (17)	Asian	1.07419	33.1019	Bladder (32)	Asian	1.24861	34.8000
Bronchus; Lung (19)	White	1.00149	31.0110	Kidney and Renal Pelvis (33)	White	1.00571	31.0647
Bronchus; Lung (19)	Black	1.00697	31.0915	Kidney and Renal Pelvis (33)	Black	1.05650	31.9157
Bronchus; Lung (19)	Asian	1.10552	32.3497	Kidney and Renal Pelvis (33)	Asian	1.27815	34.6854
Skin Melanomas (22)	White	1.00023	30.9656	Brain and Nervous System (35)	White	1.00565	31.1334
Skin Melanomas (22)	Black	1.09715	34.7281	Brain and Nervous System (35)	Black	1.02892	31.5980
Skin Melanomas (22)	Asian	1.09512	34.0309	Brain and Nervous System (35)	Asian	1.23394	34.0738

[31.0915-34.7281], 33.5291±1.0569 [31.7173-35.3975]. Figure 1 depicts the sinuosity curves, Figure 2 shows the sinuosity curves pivoted according to their respective slope angles, Figure 3 includes a box-and-whisker plot of the sinuosity indices, and Figure 4 shows a box-and-whisker plot of slope angles. The individual sinuosity indices for each cancer examined are shown in Table 1.

Discussion

The discordance of racial or ethnic groups in genomic sequencing projects potentially exacerbates racial disparities in U.S. health care, especially for complex genetic conditions. As a step towards overcoming these challenges in developing genetically informed healthcare regimens, it is of the utmost importance that the research community actively recruit racially proportional cohorts into national genomic sequencing efforts [3] such as the All of Us Research Program. Efforts such as TCGA require patient consent in the procurement of tissue samples, which implies that non-White Americans tend to withhold their samples from researchers.

The sinuosity index was adapted for this study to understand the relative differences between genomic sequence and epidemiology databases. In geography it is commonly used to study the "straightness" of an earth feature (e.g., a river). In this study, this metric was chosen as a proxy to compare the relative "straightness" of the curves that characterize the difference between the analyzed sources. To our knowledge, this is the first use of the sinuosity index in the context of biomedicine, where higher index values indicate greater dispartity. The differences between sinuosity (which quantifies shape) are more pronounced than slope alone (which measures overall difference) provide a unique persective to compare the general patterns of disparity between the races analyzed in this study.

Previous studies have shown that Black Americans are significantly less likely than their white counterparts to participate in research that used their DNA, share their DNA with a private company, and permit their DNA to be used to generate cell lines for future research [6]. Moreover, as has been noted with clinical trial studies [7], racial minorities are much less likely to want the results of their genetic testing [8] and have diverse views on the utility of genetically targeted treatments [9].

It is important to acknowledge the potential problems associated with using race as a label, especially since racial disparities in genomic sequencing may either be related to or are the direct results of the perception of how healthcare systems and research communities interest. Previous studies connecting race and variable risk for certain diseases have generated considerable controversy, as the correlation between these two factors has long been disputed among members of the scientific community and the general public.

In an effort to direct researchers away from the preconceived notion of race in genetics research, Yudell et al. noted that "the use of biological concepts of race in human genetic research - so disputed and so mired in confusion - is problematic at best and harmful at worst. It is time for biologists to find a better way." [10] In contrast, David Reich suggested that the scientific community legitimize and incorporate these claims into genetic studies while working to provide the same freedoms and opportunities to individuals irrespective of their race [11]. Reich notes that genetic studies have shown distinct differences across populations in traits such as bodily dimensions and susceptibility to diseases. Reich specifically suggests that "it will be impossible - indeed, anti-scientific, foolish, and absurd - to deny the differences [between populations]".

In contemporary medicine, treatment for cancer is usually based on the type, the size, and whether it has metastasized. As our results suggest, however, patients are at higher risk for different cancers based on their race. The results of this study further underscore that the promise of precision medicine must accomodate attributes such as race alongside pre-existing chronic conditions and environmental factors.

This study reinforces the growing wave of support for precision medicine by the government and academic institutions, specifically the United States National Institute for Minority Health and Health Disparities's aim to understand disease mechanisms that lead to differential health outcomes in minorities. This study suggests that the genetic component associated with race that affects the kinds of cancer that patients are at risk for.

For results of initiatives like the All of Us Research Program to have practical meaning to the general population, the scientific community must intently ensure that large-scale genomic sequencing efforts are representative of the range of racial backgrounds. For example, the adequate representation of female Black Americans relative to their male counterparts in TCGA indicates that these recruitment efforts into genomic sequencing studies should especially target male Black Americans. For prostate cancer, the sinuosity indices for White Americans, Black Americans, and Asian Americans were 1.01355, 1.08430, and 1.23926, indicating that Black Americans were more underrepresented in TCGA compared to White Americans (Figure 5). Interestingly, prostate cancer mortality for Black Americans is more than twice the rate observed in White Americans [5]. In contrast, Asian Americans showed overrepresentation in genomic sequencing studies, leading to disproportionately high sinuosity indices across examined cancers. This finding suggests that recruting efforts to target Asian Americans in genomic sequencing studies may not be nearly as vital as those seeking Black Americans.

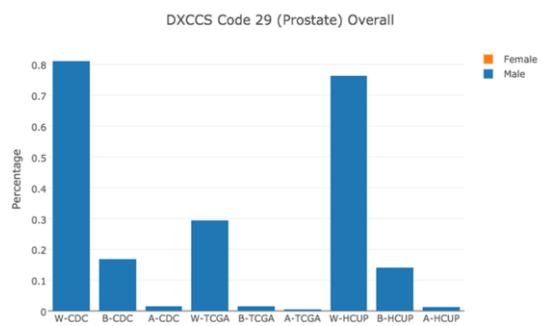

Figure 5 – Percent Representations in CDC Mortality, TCGA, and HCUP By Race for Prostate Cancer. Percent representations of prostate cancer are shown for each race (White [W], Black [B], and Asian [A]) examined across the datasets: CDC, TCGA, and HCUP.

A notable limitation of this study was being unable to examine representation for other races and some cancers due to a lack of sufficient data in at least two of the following datasets: CDC, TCGA, HCUP, and SEER. Such cancers included those of the other GI organs and peritoneum (18) and other urinary

organs (34). Thus, future work should also entail assessing adequate representation of races besides White Americans, Black Americans, and Asian Americans as well as these cancers. Relatedly, there was little to no documentation on how race data was collected in each database. This absence of documentation is relevant in that Mersha et al. found that self-reporting African Americans can have drastically different levels of African or European ancestry. Moreover, genetic analysis of individual ancestry revealed that some self-identified African Americans have significant European ancestry and vice versa [12]. Therefore, that race might be self-reported by patients in these databases would suggest that racial disparities in healthcare are largely due to factors other than genetics. Further work should be done on determining the socioeconomic and genetic basis for why race seems to play an important factor in the onset of different types of cancer. Additionally, data analysis was limited to one genomic database, so future work should incorporate genomic databases other than TCGA.

The challenge in ensuring diversity in large scale initiatives has been acknowledged. For instance, the 1000 Genomes Project analyzed the genomes of 1,092 individuals from 14 populations ranging from people with African ancestry in Southwest United States to Han Chinese in Beijing, China to British from England and Scotland. Given the importance of racial makeup in patient treatment and health outcomes, all genomic sequencing studies should follow suit to initiatives such as the 1000 Genomes Project and contain racial diversity [6]. However, it is important to note that as initiatives like All of Us launch into recruitment efforts, diversity alone is not sufficient. Diversity must be complemented with ensuring that it is with comparable frequencies relative to actual population. Otherwise, the research and clinical community risk the challenge of arriving at putative treatments that are of little utility to significant portions of the population who would benefit the most from personalized medicine approaches.

This study shows that, while race is acknowledged as an important component to consider risk for cancer, cohort studies to date are disproportionate relative to the actual occurrence of cancer by race. Thus, while genomic sequencing studies have led to novel discoveries of disease progression, they may inevitably biased by the races represented in these studies. As our results show, there has been a shortage of minorities in genomic sequencing studies, thereby potentially exacerbating racial disparities in health outcomes. Minimally, from findings of this study, we conclude that researchers must actively recruit Black Americans in genomic sequencing efforts.

The development of personalized treatments for diseases is increasingly plausible due to the increased availability of genomic data. For such efforts to be impactful, it is essential that they reflect the population of individuals that may be affected by a given disease. Despite claims of racial disparities in research populations, there have been no direct studies on such disparities in disease incidence and research projects that involve genomic sequencing. As a result, the relationship between the underrepresentation of certain races in genomic sequencing studies and health outcomes relative to these races is thus unknown.

Conclusions

This study explored the disparities in racial representation of national datasets pertaining to clinical data, mortality rates, and of a major initiative involving genomic sequence analysis (The Cancer Genome Atlas [TCGA]). The results suggest that Black Americans are underrepresented for most cancers in TCGA compared to clinical and mortality datasets, whereas Asian Americans are overrepresented. Additionally, male Black Americans tend to be especially underrepresented in such genomic sequencing studies compared to their female counterparts. These findings highlight the importance of targeted efforts to actively recruit representative patient populations into studies involving genomic sequencing.

Acknowledgements

This study was funded by grant U54GM115467 from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] S.J. Aronson, H.L. Rehm, Building the foundation for genomics in precision medicine, *Nature* 526 (2015), 336-342.
[2] Y.S. Ha, A. Salmasi, M. Karellas, E.A. Singer, J.H. Kim, M. Han, A.W. Partin, W.J. Kim, D.H. Lee, and I.Y. Kim, Increased incidence of pathologically nonorgan confined prostate cancer in African-American men eligible for active surveillance, *Urology* 81 (2013), 831-835.
[3] L.G. Landry, N. Ali, D.R. Williams, H.L. Rehm, and V.L. Bonham, Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice, *Health Aff* 37 (2018), 780-785.
[4] D.E. Spratt, T. Chan, L. Waldron, C. Speers, F.Y. Feng, O.O. Ogunwobi, and J.R. Osborne, Racial/Ethnic disparities in genomic sequencing, *JAMA Oncol* 2 (2016), 1070-1074.
[5] E.D. Lazarus and J.A. Constantine, Generic theory for channel sinuosity, *Proc. Natl. Acad. Sci. U.S.A.* 110 (2013), 8447-8452.
[6] T. Dye, D. Li, M. Demment, S. Groth, D. Fernandez, A. Dozier, and J. Chang, Sociocultural variation in attitudes toward use of genetic information and participation in genetic research by race in the United States: implications for precision medicine, *J. Am. Med. Inform. Assoc.* 23 (2016), 782-786.
[7] K.L. Pariera, S.T. Murphy, J. Meng, and M.L. McLaughlin, Exploring willingness to participate in clinical trials by ethnicity, *J Racial Ethn Health Disparities* 4 (2017), 763-769.
[8] R.G. Salloum, T.J. George, N. Silver, M.J. Markham, J.M. Hall, Y. Guo, J. Bian, and E.A. Shenkman, Rural-urban and racial-ethnic differences in awareness of direct-to-consumer genetic testing, *BMC Public Health* 18 (2018), 277.
[9] C.H. Halbert, J.A. McDonald, G. Magwood, and M. Jefferson, Beliefs about genetically targeted care in African Americans, *J. Natl. Med. Assoc.* 109 (2017), 98-106.
[10] M. Yudell, D. Roberts, R. DeSalle, and S. Tishkoff, Taking race out of human genetics, *Science* 351 (2016), 564-565.
[11] D. Reich, How genetics is changing our understanding of 'race', *The New York Times* (2018).
[12] T.B. Mersha and T. Abebe, Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities, *Hum Genomics* 9 (2015): 1-15.

Address for correspondence

Indra Neil Sarkar, PhD, MLIS, FACMI
Center for Biomedical Informatics
Brown University
Box G-R
Providence, RI 02912
neil_sarkar@brown.edu