Brief Report

Racial/Ethnic Disparities in Genomic Sequencing

Daniel E. Spratt, MD; Tiffany Chan, MA; Levi Waldron, PhD; Corey Speers, MD; Felix Y. Feng, MD; Olorunseun O. Ogunwobi, MD, PhD; Joseph R. Osborne, MD, PhD

IMPORTANCE Although poorly understood, there is heterogeneity in the molecular biology of cancer across race and ethnicities. The representation of racial minorities in large genomic sequencing efforts is unclear, and could have an impact on health care disparities.

OBJECTIVE To determine the racial distribution among samples sequenced within The Cancer Genome Atlas (TCGA) and the deficit of samples needed to detect moderately common mutational frequencies in racial minorities.

DESIGN, SETTING, AND PARTICIPANTS This was a retrospective review of individual patient data from TCGA data portal accessed in July 2015. TCGA comprises samples from a wide array of institutions primarily across the United States. Samples from 10 of the 31 currently available tumor types were analyzed, comprising 5729 samples from the approximately 11 000 available.

MAIN OUTCOMES AND MEASURES Using the estimated median somatic mutational frequency, the samples needed beyond TCGA to detect a 10% and 5% mutational frequency over the background somatic mutation frequency were calculated for each tumor type by racial ethnicity.

RESULTS Of the 5729 samples, 77% (n = 4389) were white, 12% (n = 660) were black, 3% (n = 173) were Asian, 3% (n = 149) were Hispanic, and less than 0.5% combined were from patients of Native Hawaiian, Pacific Islander, Alaskan Native, or American Indian decent. This overrepresents white patients compared with the US population and underrepresents primarily Asian and Hispanic patients. With a somatic mutational frequency of 0.7 (prostate cancer) to 9.9 (lung squamous cell cancer), all tumor types from white patients contained enough samples to detect a 10% mutational frequency. This is in contrast to all other racial ethnicities, for which group-specific mutations with 10% frequency would be detectable only for black patients with breast cancer. Group-specific mutations with 5% frequency would be undetectable in any racial minority, but detectable in white patients for all cancer types except lung (adenocarcinoma and squamous cell carcinoma) and colon cancer.

CONCLUSIONS AND RELEVANCE It is probable, but poorly understood, that ethnic diversity is related to the pathogenesis of cancer, and may have an impact on the generalizability of findings from TCGA to racial minorities. Despite the important benefits that continue to be gained from genomic sequencing, dedicated efforts are needed to avoid widening the already pervasive gap in health care disparities.

JAMA Oncol. doi:10.1001/jamaoncol.2016.1854
Published online June 30, 2016.

Author Affiliations: Department of Radiation Oncology, University of Michigan, Ann Arbor (Spratt, Speers, Feng); School of Public Health, City University of New York, Hunter College, New York (Chan, Waldron); Department of Biological Sciences, City University of New York, Hunter College, New York (Ogunwobi); Department of Medicine, Cornell University, Weill Cornell Medicine, New York (Ogunwobi); Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, New York (Osborne).

Corresponding Author: Joseph R. Osborne, MD, PhD, Molecular Imaging and Therapy Service, Department of Radiology, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY 10065 (osbornej@mskcc.org).

Two of the 27 Institutes and Centers of the National Institutes of Health of the US Department of Health and Human Services, namely the National Cancer Institute and the National Human Genome Research Institute, have teamed together to support the creation of The Cancer Genome Atlas (TCGA), a series of cross-sectional, comprehensive genomic studies of more than 11 000 patients with 31 cancer types collected to date. The cohort composition for each disease site is of critical importance because these sites are intended to represent the respective disease among the general population. However, it is probable, but poorly understood, that racial diversity is intimately related to the pathogenesis of cancer, and may have an impact on the generalizability of findings from these data sets.[1]

A prototypic example of racial diversity among the mutational landscape of cancer is the high prevalence of *EGFR* mutations among patients of Asian descent (estimated to occur in approximately 50% of the Asian population).[2] The ability to confidently detect mutations in a particular subgroup of patients depends on the background mutational frequency (ie, noise), the mutational rate of the target of interest (ie, signal), and the *absolute* sample size (ie, number of tumors sequenced). Sufficiently large sample sizes are necessary to provide power to detect infrequent mutations confidently over the background rate.[3] However, the mutational frequency we are able to detect in racial minorities among large sequencing efforts, such as TCGA, is currently unknown. TCGA project has uncovered numerous uncommon subtypes and mutations across multiple cancer types, and these results are being used to develop new therapies and ultimately improve outcomes for patients with cancer. However, without adequate representation of racial minorities within massive sequencing efforts, health care disparities may inadvertently be increased because race-specific mutational patterns are unable to be appreciated.[4]

Key Points

Question What is the racial distribution among samples sequenced within The Cancer Genome Atlas and the deficit of samples needed to detect moderately common mutational frequencies in racial minorities?

Findings A review of individual patient data from 5729 samples showed that only 12% were black, 3% were Asian, and 3% were Hispanic. For no racial minorities could we detect a mutational frequency of 5% in any cancer type analyzed.

Meaning There are insufficient samples from racial minorities to detect moderately common genomic alterations in this population, which may be inadvertently widening the already pervasive gap in healthcare disparities.

Methods

Using TCGA data portal accessed in July 2015, clinical and level 3 mutational data were collected from 10 of the 31 available tumor types: breast, prostate, lung adenocarcinoma, lung squamous cell carcinoma (SCC), colon, renal clear cell, uterine, ovarian, head and neck SCC, and glioblastoma multiforme.

Demographic data were extracted and merged from level 1 and level 4 data, including categories of race, ethnicity, age, and sex. The categories used are presented in the **Table**, and are as defined in the TCGA data set; the terms *race* and *ethnicity* were not defined by the authors and were used per TCGA data fields. Racial categories included white, black or African American, Asian, Native Hawaiian or Pacific Islander, and American Indian or Alaskan Native. Ethnic categories included Hispanic and non-Hispanic. Samples without racial or ethnic information were recorded as well.

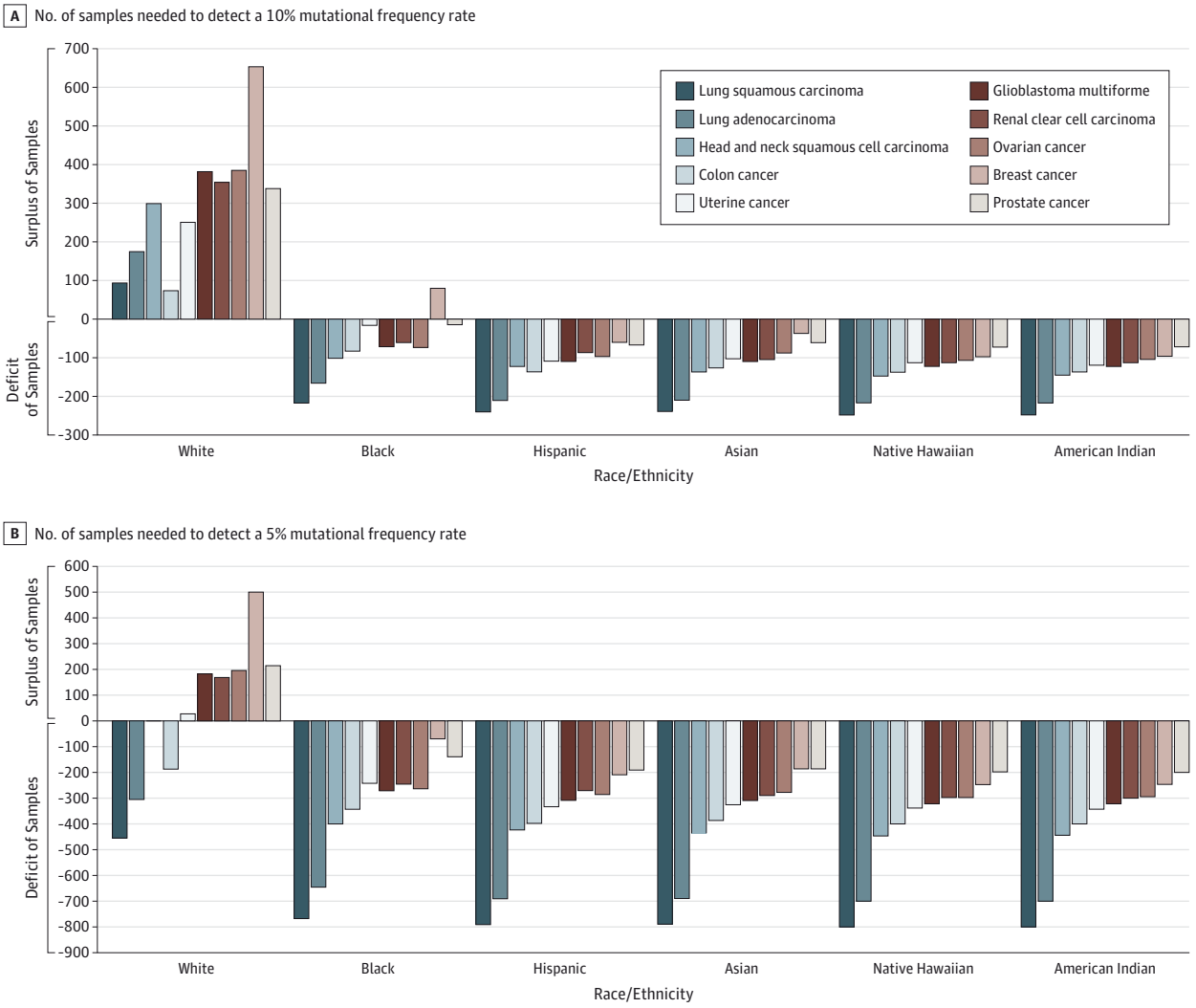
Table. Demographic Characteristics of Common Cancers in TCGA

Characteristic	Cancer, %										Total, No. (%)
	Prostate	Breast	Lung AC	Lung SCC	Colon	Uterine	Ovarian	HNSCC	Kidney	GBM	
Total No.	495	1085	521	495	453	540	591	423	533	593	5729
Race[a]											
White	83	69	75	69	47	69	83	106	87	85	4389 (77)
Black	12	16	10	6	12	19	6	11	10	9	660 (12)
Asian	2	6	2	2	2	4	3	3	2	2	173 (3)
Native Hawaiian	0	0	0	0	0	2	0	0	0	0	10 (0.2)
American Indian	0	0	0	0	0	1	1	0	0	0	13 (0.2)
NA	3	9	13	23	38	5	7	4	1	4	584 (10)
Ethnicity[a]											
Hispanic	1	4	1	2	0	3	2	6	5	2	149 (3)
Non-Hispanic	78	80	74	63	59	69	57	109	66	82	4234 (74)
NA	21	16	24	36	41	28	41	9	29	16	1446 (25)
Median somatic mutation frequency (per Mb)	0.7	1.2	8.1	9.9	3.1	2.5	1.7	3.9	1.9	2.2	

Abbreviations: AC, adenocarcinoma; GBM, glioblastoma multiforme; HNSCC, squamous cell carcinoma; NA, not available; SCC, squamous cell carcinoma; TCGA, The Cancer Genome Atlas.

[a] The terms "race" and "ethnicity" were the terminology used in the TCGA data sets "or African American," "or Pacific Islander," "or Alaskan Native."

Figure. Numbers of Samples by Race/Ethnicity Needed to Detect 10% and 5% Mutational Frequencies Above 10 Cancers' Background Mutational Frequency Rate Sequenced by TCGA



A, Numbers to detect a 10% mutational frequency. B, Numbers to detect a 10% mutational frequency. TCGA indicates The Cancer Genome Atlas.

The median somatic mutation frequency (per Mb) of each cancer has been previously reported.[3] Briefly, the power to determine if a gene is significantly mutated depends on the target mutation frequency above background and the average background somatic mutation frequency of the cancer type. Using these data, we estimated the sample size needed to detect a 10% and 5% mutational frequency over the somatic mutational frequency rate with 90% power in 90% of genes. The available sample size from each racial group within TCGA was subtracted from this calculated sample size to determine either the surplus of deficit of samples needed to detect the respective mutational frequency rate.

Results

Of the 5729 samples, 77% (n = 4389) were white, 12% (n = 660) were black, 3% (n = 173) were Asian, 3% (n = 149) were Hispanic, and less than 0.5% combined were from patients of Native Hawaiian, Pacific Islander, Alaskan Native, or American Indian descent (Table). This is in comparison to the US population demography: 64% white, 12% black, 5% Asian, 16% Hispanic, 1% to 2% Native Hawaiian, Pacific Islander, Alaskan Native, or American Indian decent. This overrepresents white patients compared with the US population and underrepresents primarily Asian and Hispanic patients.

With somatic mutational frequencies of 0.7 (prostate cancer) to 9.9 (lung SCC) (Table), all tumor types from white patients contained enough samples to detect a 10% mutational frequency (**Figure**, A). This is in contrast to all other races/ethnicities, for which adequate sample size to detect the same mutational frequency existed only for black patients with breast cancer. In no cancer type would in any racial minority would a mutational frequency of 5% be detectable, whereas a 5% mutational frequency could be detected in all tumor types

of white patients except lung (adenocarcinoma and SCC) and colon cancer (Figure, B).

Discussion

As we demonstrate, despite approximately proportional *relative* sample size of many demographic minorities within TCGA when compared with the US population, the *absolute* sample size of these minorities is inadequate to capture even relatively common somatic mutations that are specific to those groups. Still, TCGA can be commended for their enrollment of racial minorities that has been far more successful than many clinical trial efforts.[5]

Importantly, one of the fastest-growing patient populations in the United States is of Asian descent. However, our data suggest that they are significantly underrepresented in TCGA (approximately 66% underrepresented). Interestingly, the best-known example of a targetable mutation in cancer that varies by race/ethnicity is arguably the *EGFR* mutation in lung adenocarcinoma. The phase 3 randomized clinical trial Iressa Survival Evaluation in Advanced Lung Cancer (ISEL) failed to demonstrate a benefit of using gefitinib, a small-molecule inhibitor of *EGFR* in all-comers in a predominantly white cohort.[6] However, a preplanned subgroup analysis showed a significant overall survival benefit in Asian patients. These observations are explained by the PIONEER study, a multinational epidemiologic prospective study that demonstrated that *EGFR* mutations are present in 51.4% of stage IIIB or IV lung adenocarcinomas among Asian patients, in contrast to approximately 20% in white and African American patients.[2] Given the potential for disparate tumor biology by race, we must critically evaluate the generalizability of new discoveries to all patients.

Not all mutations or genomic alterations are as common as *EGFR* mutations in non–small-cell lung cancer. Another recent success in targeted therapy is targeting the relatively infrequent genomic alteration of *ALK* rearrangement in non–small-cell lung cancer (approximately 4% in unselected patients).[7] Other examples from large genomic analyses of lung cancer include *BRAF* mutations, which in 1 study[8] occurred in 3% (18 of 697) of patients, all of whom were from white patients. In racial minorities, there may be undiscovered low-frequency mutations that could also result in the use of new targeted therapies.

Increasing the representation of racial minorities will also enable analyses to determine what drives aggressive tumor biology across races/ethnicities. As we have demonstrated, black women with breast cancer were the only subset to have ample representation of black patients to detect a less than 10% mutational frequency rate over background. This opportunity has led to novel data demonstrating that this group has greater intra-tumor heterogeneity and basal gene-expressing tumors by about 2-fold compared with white patients.[9]

The burden of this problem should not rest on TCGA, and a key to overcoming the lack of minority participation in sequencing efforts is the sharing of clinical and genomic data across institutions, academia, and industry. An example of this was performed by Yamoah et al,[10] who acquired approximately 3 times the number of black prostate cancer samples compared with TCGA, and identified a potential ethnicity-dependent biomarker to predict prostate cancer outcomes. Furthermore, multinational efforts will also be critical to determine if there are differences in racially biased mutations in endemic and nonendemic areas despite similar racial ancestry.

Limitations of this study exist. Only 10 cancer types of TCGA were investigated, and other large sequencing efforts were not investigated. A relatively large percentage of patients in TCGA had missing racial and/or ethnicity information, which may alter our findings.

Conclusions

Low absolute enrollment of minority patients in cancer sequencing studies limits the ability to detect targetable mutations specific to minority groups. Even proportional enrollment of minorities could have lasting implications on disparities in treatment and outcome, and amplify existing inequalities in health care delivery and patient outcomes.

ARTICLE INFORMATION

Accepted for Publication: April 5, 2016.

Published Online: June 30, 2016. doi:10.1001/jamaoncol.2016.1854.

Author Contributions: Dr Spratt had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Dr Spratt and Ms Chang contributed equally to this study.
Study concept and design: All authors.
Acquisition, analysis, or interpretation of data: All authors.
Study concept and design: Spratt, Chan, Ogunwobi, Osborne.
Acquisition, analysis, or interpretation of data: Spratt, Chan, Waldron, Speers, Feng, Ogunwobi.
Drafting of the manuscript: Spratt, Chan, Waldron, Ogunwobi.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Spratt, Chan, Waldron.
Obtained funding: Spratt.
Administrative, technical, or material support: Feng, Ogunwobi, Osborne.
Study supervision: Spratt, Waldron, Speers, Ogunwobi, Osborne.

Conflict of Interest Disclosures: Dr Feng serves on the advisory boards of Medivation/Astellas, GenomeDx, Nanostring, and Celgene. No other disclosures are reported.

Funding/Support: Dr Osborne is supported by U54 CA137788 (CCNY-MSKCC Partnership for Cancer Research, Training, and Community Outreach) and R21 CA153177-03 Center to Reduce Cancer Health Disparity (principal investigator, Dr Osborne). Drs Spratt and Feng receive funding from the Prostate Cancer Foundation.

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Calvo E, Baselga J. Ethnic differences in response to epidermal growth factor receptor tyrosine kinase inhibitors. *J Clin Oncol*. 2006;24(14):2158-2163.

2. Shi Y, Au JS-K, Thongprasert S, et al. A prospective, molecular epidemiology study of EGFR mutations in Asian patients with advanced non-small-cell lung cancer of adenocarcinoma histology (PIONEER). *J Thorac Oncol*. 2014;9(2):154-162.

3. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505(7484):495-501.

4. Ateeq B, Bhatia V, Goel S. Molecular discriminators of racial disparities in prostate cancer. *Trends in Cancer*. 2016:2(3);116-119. doi:10.1016/j.trecan.2016.01.005.

5. Spratt DE, Osborne JR. Disparities in castration-resistant prostate cancer trials. *J Clin Oncol*. 2015;33(10):1101-1103.

6. Thatcher N, Chang A, Parikh P, et al. Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer). *Lancet*. 2005;366(9496):1527-1537.

7. Shaw AT, Kim D-W, Nakagawa K, et al. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N Engl J Med*. 2013;368(25):2385-2394.

8. El-Telbany A, Ma PC. Cancer genes in lung cancer: racial disparities: are there any? *Genes Cancer*. 2012;3(7-8):467-480.

9. Keenan T, Moy B, Mroz EA, et al. Comparison of the genomic landscape between primary breast cancer in African American vs white women and the association of racial differences with tumor recurrence. *J Clin Oncol*. 2015;33(31):3621-3627.

10. Yamoah K, Johnson MH, Choeurng V, et al. Novel biomarker signature that may predict aggressive disease in African American men with prostate cancer. *J Clin Oncol*. 2015;33(25):2789-2796.