Computational and Structural Biotechnology Journal 21 (2023) 5413–5422

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Pan-cancer mutational signature surveys correlated mutational signature with geospatial environmental exposures and viral infections

Judy Bai [a], Katherine Ma [a], Shangyang Xia [a], Richard Geng [a], Claire Shen [a], Limin Jiang [a], Xi Gong [c], Hui Yu [a], Shuguang Leng [b], Yan Guo [a,*]

[a] Department of Public Health and Sciences, Sylvester Comprehensive Cancer Center, University of Miami, FL 33136, USA
[b] Comprehensive Cancer Center, Albuquerque, University of New Mexico, NM 87109, USA
[c] Geography & Environmental Studies, University of New Mexico, Albuquerque, NM 87109, USA

ARTICLE INFO

Keywords:
Cancer
Environmental pollution
Mutational signature
Racial disparity
Oncovirus

ABSTRACT

Background: Cancer has been disproportionally affecting minorities. Genomic-based cancer disparity analyses have been less common than conventional epidemiological studies. In the past decade, mutational signatures have been established as characteristic footprints of endogenous or exogenous carcinogens.

Methods: Integrating datasets of diverse cancer types from The Cancer Genome Atlas and geospatial environmental risks of the registry hospitals from the United States Environmental Protection Agency, we explored mutational signatures from the aspect of racial disparity concerning pollutant exposures. The raw geospatial environmental exposure data were refined to 449 air pollutants archived and modeled from 2007 to 2017 and aggregated to the census county level. Additionally, hepatitis B and C viruses and human papillomavirus infection statuses were incorporated into analyses for skin cancer, cervical cancer, and liver cancer.

Results: Mutation frequencies of key oncogenic genes varied substantially between different races. These differences were further translated into differences in mutational signatures. Survival analysis revealed that the increased pollution level is associated with worse survival. The analysis of the oncogenic virus revealed that aflatoxin, an affirmed carcinogen for liver cancer, was higher in Asian liver cancer patients than in White patients. The aflatoxin mutational signature was exacerbated by hepatitis infection for Asian patients but not for White patients, suggesting a predisposed genetic or genomic disadvantage for Asians concerning aflatoxin.

Conclusions: Environmental pollutant exposures increase a mutational signature level and worsen cancer prognosis, presenting a definite adverse risk factor for cancer patients.

1. Introduction

As a leading cause of death, cancer receives ongoing, timely, and systematic surveillance at national and worldwide strategic levels [1]. The incidence, mortality, and survival of a specific cancer type vary geographically and differ across ethnicity and population boundaries, giving rise to the issue of cancer disparity. Cancer disparities must be closely monitored, analyzed, and, ideally, suppressed. In the U.S., cancer disparity is a complex issue attributed to a series of intertwined factors, including poverty, lifestyle, education, income, environmental exposures, and others [2]. Fortunately, the gravity of cancer disparity has garnered substantial research attention, and substantial efforts have been put into epidemiological analyses of ethnic and environmental risk factors associated with cancer disparity. Often, the effects of cultural or

Abbreviations: ACC, Adrenocortical Carcinoma; BLCA, Bladder Urothelial Carcinoma; BRCA, Breast Invasive Carcinoma; CESC, Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma; CHOL, Cholangiocarcinoma; COAD, Colon Adenocarcinoma; DLBC, Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; ESCA, Esophageal Carcinoma; GBM, Glioblastoma Multiforme; HNSC, Head and Neck Squamous Cell Carcinoma; KICH, Kidney Chromophobe; KIRC, Kidney Renal Clear Cell Carcinoma; KIRP, Kidney Renal Papillary Cell Carcinoma; LAML, Acute Myeloid Leukemia; LGG, Brain Lower Grade Glioma; LIHC, Liver Hepatocellular Carcinoma; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Cell Carcinoma; MESO, Mesothelioma; OV, Ovarian Serous Cystadenocarcinoma; PAAD, Pancreatic Adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PRAD, Prostate Adenocarcinoma; READ, Rectum Adenocarcinoma; SARC, Sarcoma; SKCM, Skin Cutaneous Melanoma; STAD, Stomach Adenocarcinoma; TGCT, Testicular Germ Cell Tumors; THCA, Thyroid Carcinoma; THYM, Thymoma; UCEC, Uterine Corpus Endometrial Carcinoma; UCS, Uterine Carcinosarcoma; UVM, Uveal Melanoma.

* Corresponding author.
E-mail address: yanguo1978@gmail.com (Y. Guo).

https://doi.org/10.1016/j.csbj.2023.10.041
Received 25 July 2023; Received in revised form 18 October 2023; Accepted 19 October 2023
Available online 20 October 2023

socioeconomic inequalities are explored in extra depth. For example, using Whites as a reference, Ellis et al. found Black cancer patients had the lowest survival while Asians had the highest survival [3]. The stage at diagnosis was the largest attributing factor for the survival disparity, which can be directly tied to the socioeconomic status of each racial group. In another study, Butler et al. approached cancer disparity from the unique angle of physician cultural competency [4], finding that minorities were not adequately matched with culturally competent physicians, and this suboptimal reality might have impacted the overall quality of care for minorities.

In the U.S., it is well documented that minorities are disproportionally affected by environmental pollution. For example, it was found that racial disparity in pollution exposure is strongest among neighborhoods with median incomes below $25,000, which predominantly represent Blacks and Hispanics [5]. Another recent study shows that there is a 45-fold difference in average pollution exposure between the most and least exposed, disproportionally affecting Blacks the most [6]. Liu et al. examined the disparities between race and ethnicity in air pollution exposure across critical air pollutants within the U.S. between 1990 and 2010, and found that racial and ethnic minorities were more than twice as likely as non-Hispanic Whites to live in a census block group with the highest levels of air pollution [7]. There is also evidence in other national studies that non-Hispanic Black populations were more likely to live closer to industrial $PM_{2.5}$ emissions [8] and experience higher average exposures to industrial air toxins [9] than other race and ethnicity groups. According to Bell and Ebisu's study conducted across 215 U.S. census tracts in 2000–2006, Whites generally were exposed to $PM_{2.5}$ components at the lowest level and Hispanics at the highest level; non-Hispanic Blacks were exposed to 13 of the 14 $PM_{2.5}$ components at higher levels than Whites [10]. Kravitz-Wirtz et al. reported that environmental inequality can be observed at the level of census blocks, with non-Hispanic Blacks and Hispanic residents having significantly higher levels of NO_2, $PM_{2.5}$, and PM_{10} exposures than White residents [11].

In the past two decades, high-throughput sequencing platforms have been continually optimized, along with steadily decreasing costs. As a result, genome-wide variant analysis becomes a promising approach to cancer research. The accumulated cancer variant data thus allow for examination of cancer differences from a genomics perspective. Among all directions around genomic variants, an emerging, prospective approach to carcinogenesis is through the paradigm of "mutational signatures", with a mutational signature representing a characteristic combination of somatic variant types and theoretically mapping to a specific cancer etiology [12–14]. Here, we focus on mutational signatures defined around single base substitution (SBS). To date, there are nearly 100 well-defined mutational signatures [14], and more than half are associated with known etiologies. For example, mutational signature SBS7 represents mutation footprints caused by UV light exposure, SBS4 represents tobacco smoking, and SBS24 represents aflatoxin exposure. Because a mutational signature is presumed to be an aggregate of all types of SBS somatic variants, they may be used as quantitative intermediate surrogate phenotypes for otherwise unattainable environmental variables, enabling delving into cancer racial variance at the etiology level. In literature, artificial exposure to carcinogens has been proven to induce mutational signatures [15]. On the contrary, no empirical evidence has clearly linked environmental pollutants to mutational signatures based on human data.

Conventionally, cancer disparity studies are set against cohort-level epidemiological measures such as incidence, prevalence, mortality, survival, mortality, and morbidity. With the advancements in measurement and interpretation of genomics data, burgeoning studies are able to take summary statistics for an individual genome, epigenome, or transcriptome, such as mutational burden [16,17], complex arm aberration index [18], and Alu editing index [19]. In this study, we envisioned that similar summary indices built around individual mutational signatures may form proxies for intermediate phenotypes indicative of environmental toxic exposures. Additionally, decomposing genome-wide mutations into diverse mutational signatures allowed for a finer granularity to delve into the etiological heterogeneity of the same clinically defined cancer. With these principles in mind, we conducted a mutational-signature-based cancer study by incorporating SBS variants and environmental exposures (air pollutants) of tens of thousands of cancer patients into an array of statistical analyses.

2. Methods

2.1. Genomic, survival, and mutational signature data

Variant data for 10,182 patients classified to 33 cancer types were downloaded from Genomic Data Commons, the data portal of The Cancer Genome Atlas (TCGA). Disease specific survival data of the same cohorts of cancer patients were obtained from Pan-Cancer Data Resource [20]. During data analysis, we excluded datasets with race sample size less 30. The SBS mutational signature reference file (v3.3) was downloaded from The Catalog Of Somatic Mutations In Cancer [14], and we adopted 49 well-defined, non-artifact mutational signatures [21]. These reference mutational signatures are named with SBS prefixes, e.g. SBS4.

Each subject's variant data were quantified to a catalog of 96 three-base motifs centered upon the mutated SBS (upstream, SBS, downstream), and all patients' mutation catalogs for one cancer type were fitted by R package MutationalPatterns [22] to infer the level (quantitative contribution) of each reference mutational signature in each patient. The fitting function from MutationalPatterns outputted a subject's mutational signature level as a non-negative value, indicating the quantitative contribution of this particular signature to the mutation catalog of the patient. A zero value for mutational signature level means this particular signature was not present in the subject, contributing a null variant.

2.2. Air pollution data from United States environmental protection agency

According to the geographical locations of patients' visiting hospitals, the environmental exposure of cancer patients was surveyed on a panel of 449 air pollutants modeled and archived by the United States Environmental Protection Agency (EPA). The air pollution data were obtained from Risk-Screening Environmental Indicators (RSEI) modeled and archived by the United States Environmental Protection Agency (U. S. EPA). Since 1987, U.S. facilities in different industrial sectors have been required to report their environmental releases of more than 600 toxic chemicals on an annual basis under the U.S. EPA's Toxics Release Inventory (TRI) program. RSEI incorporates TRI information as well as other data sources and risk factor concepts in order to evaluate the potential impacts of industrial emissions of TRI-listed chemicals. The American Meteorological Society/EPA Regulatory Model (AERMOD), a steady-state Gaussian plume model (https://www.epa.gov/scram/air-quality-dispersion-modeling-preferred-and-recommended-models), was used in RSEI to model the dispersion of air emissions of toxic chemicals, which produces detailed air pollution concentration modeling results at various spatial and temporal scales for data users. This study collected the RSEI-modeled annual average concentrations for all available air pollutants in each grid cell of 810 m by 810 m in the United States during the period 2007–2017. In total, 449 chemicals were modeled in each of the 11 years, which were kept for further analysis. The unit of the pollutant is the concentration of chemicals at grid cell (μg/m3). Afterward, modeled concentrations were aggregated to the county level by averaging the values across all grids within the respective county. An individual TCGA subject's exposure to a particular chemical was represented by the 11-year average concentrations of the county in which the visiting hospital or reporting agency is located. Thus, each TCGA subject is associated with exposure values for each of the 449 toxic air pollutants calculated from the RSEI. The pollution data used in this study is viewable in Supplementary Table S1.

2.3. Data analysis

All statistical analyses and data science operations were performed in the open source environment R (v3.6). All analyses were performed for each cancer type in parallel. For all regression analyses, we adjusted for clinical variables including age, sex, tumor stage, and race wherever possible. The race definition was based on TCGA's clinical data. For all statistical tests, when applicable, the Benjamini-Hochberg adjusted p-value < 0.05 or adjusted p-value < 0.01 was used as the statistical significance threshold.

First, logistic regression was adopted to discern the difference in a binary variable between two (racial) groups. For this particular type of analyses, a variant in a single gene in a subject was treated as a binary variable, and the continuous level of a mutational signature in a subject was dichotomized to a binary variable of zero vs. non-zero. In brief, for each single gene or each single mutational signature, we calculated the proportion of subjects presenting a variant or a non-zero mutational signature contribution and compared the proportions in two comparative groups with logistic regression. In most analyses, patients were divided into groups by race, but at times, they were grouped by the geographical locations of the hospitals they visited. Such logistic regressions were conducted with respect to all genes and all mutational signatures in a repetitive manner.

Second, using a sensitivity analysis, we modeled the level of a mutational signature in a subject with either a univariate linear regression on race (Eq. 1) or a multivariate linear regression on race along with a pollutant (Eq. 2). Here, the level of a mutational signature, *MS*, retained its continuous nature as outputted by MutationalPattern (without dichotomization). The multivariate linear regression included race and pollutant as independent variables: race was coded as a categorical variable (*R*), and pollutant was a continuous value representing the intensity of an air pollutant (*P*). The other notations of the equations referred to the constant term (*a*) and the coefficients of the independent variables (*b*, *b1*, and *b2*). Each subject contributed a sample or an instantiation of the regression model. For every pollutant, the model of inclusion (Eq. 2) and the corresponding model of exclusion (Eq. 1) were compared via the ANOVA test, so that the net contribution of the pollutant variable was assessed in addition to the race effect.

Third, logistic regression was deployed in addition to linear regression to mitigate vulnerability to extreme levels of mutational signatures. In addition to the aforementioned linear regression analyses (Eqs. 1 and 2) , we employed logistic regression in similarly parallel frameworks after dichotomizing mutational signature intensities (*MS*) to zero and none-zero values (*MS'*, Eqs. 3 and 4). Again, ANOVA tests were performed to assess the net contribution of the additional pollutant variable. A pollutant was considered as significantly associated with a mutational signature only if all four adjusted p-values were significant: the two p-values for the pollutant coefficient *b2* from both the linear and the logistic regressions (Eqs. 2 and 4), and the two ANOVA p-values for the Eq. 1 vs. Eq. 2 and Eq. 3 vs. Eq. 4 comparisons.

$$MS = a + b \bullet R \tag{1}$$

$$MS = a + b_1 \bullet R + b_2 \bullet P \tag{2}$$

$$Prob(MS') = \frac{e^{(a+b\bullet R)}}{1 + e^{(a+b\bullet R)}} \tag{3}$$

$$Prob(MS') = \frac{e^{(a+b1\bullet R+b2\bullet P)}}{1 + e^{(a+b1\bullet R+b2\bullet P)}} \tag{4}$$

Furthermore, with respect to the disease-specific survival data, we implemented Cox Proportional Hazard models to investigate if the survival span showed a difference for pollution at high and low levels. The survival analysis was only applied to datasets with the number of event

Table 1
Genes that displayed significant differences in variant frequency between a minority group and Whites.

Cancer	Gene	Freq1[a]	Freq2[b]	Adjusted P[c]	Minority[d]
BRCA	TP53	28.5 %	42.3 %	0.005	Black
BRCA	PYHIN1	0.6 %	8.5 %	0.005	Asian
BRCA	ITGAM	0.4 %	8.5 %	0.006	Asian
BRCA	FBXW7	1.0 %	5.5 %	0.008	Black
BRCA	PRPF8	0.4 %	6.8 %	0.011	Asian
BRCA	FOXO4	0.4 %	6.8 %	0.011	Asian
BRCA	PPP6R3	0.4 %	6.8 %	0.011	Asian
BRCA	SSX2IP	0.3 %	5.1 %	0.012	Asian
BRCA	OR10J3	0.3 %	5.1 %	0.013	Asian
BRCA	CENPE	1.8 %	8.5 %	0.013	Asian
BRCA	PHC2	0.4 %	5.1 %	0.013	Asian
BRCA	EYS	1.8 %	8.5 %	0.015	Asian
BRCA	ZNF334	0.6 %	5.1 %	0.017	Asian
BRCA	SCN8A	1.1 %	6.8 %	0.020	Asian
BRCA	HERC6	0.9 %	6.8 %	0.020	Asian
BRCA	MRPL2	0.4 %	5.1 %	0.020	Asian
BRCA	CDCP1	0.4 %	5.1 %	0.020	Asian
BRCA	ZNF776	0.4 %	5.1 %	0.020	Asian
BRCA	ZNF252P	0.1 %	5.1 %	0.020	Asian
BRCA	IFT88	0.3 %	5.1 %	0.021	Asian
BRCA	OPLAH	0.4 %	5.1 %	0.022	Asian
BRCA	MCM3	0.1 %	5.1 %	0.022	Asian
BRCA	TP53	28.5 %	49.2 %	0.022	Asian
BRCA	CPXM1	0.4 %	5.1 %	0.023	Asian
BRCA	QSER1	1.4 %	6.8 %	0.023	Asian
BRCA	KMT2D	2.4 %	8.5 %	0.023	Asian
BRCA	TRIM42	0.4 %	5.1 %	0.023	Asian
BRCA	FTSJ3	1.0 %	5.1 %	0.023	Asian
BRCA	MSL2	0.3 %	5.1 %	0.023	Asian
BRCA	SLCO5A1	0.7 %	5.1 %	0.024	Asian
BRCA	PREX2	2.0 %	8.5 %	0.024	Asian
BRCA	MAP10	0.6 %	5.1 %	0.024	Asian
BRCA	GOLGB1	1.6 %	6.8 %	0.024	Asian
BRCA	CSTF3	0.9 %	5.1 %	0.024	Asian
BRCA	ABCA5	0.9 %	5.1 %	0.024	Asian
BRCA	EFEMP1	0.4 %	5.1 %	0.024	Asian
BRCA	PDCL3	0.4 %	5.1 %	0.024	Asian
KIRC	VHL	48.5 %	20.8 %	0.026	Black
BRCA	SMARCC2	0.6 %	5.1 %	0.028	Asian
KIRC	CASP8AP2	0.7 %	7.5 %	0.030	Black
BRCA	POTEA	0.6 %	5.1 %	0.030	Asian
BRCA	ARID1A	2.6 %	10.2 %	0.030	Asian
BRCA	CYP11B2	0.6 %	5.1 %	0.030	Asian
BRCA	CSPP1	1.1 %	5.1 %	0.031	Asian
BRCA	MROH5	0.4 %	5.1 %	0.031	Asian
BRCA	TARBP1	0.6 %	5.1 %	0.033	Asian
BRCA	SLIT2	1.4 %	6.8 %	0.033	Asian
BRCA	MYH4	1.1 %	6.8 %	0.034	Asian
BRCA	TRPC5	1.1 %	5.1 %	0.036	Asian
BRCA	SSH2	1.0 %	5.1 %	0.036	Asian
BRCA	LAMB4	1.6 %	6.8 %	0.037	Asian
BRCA	PTPRB	2.0 %	6.8 %	0.038	Asian
BRCA	MADD	1.1 %	5.1 %	0.038	Asian
BRCA	MICAL2	0.7 %	5.1 %	0.038	Asian
BRCA	SPINK5	0.7 %	5.1 %	0.038	Asian
BRCA	UACA	0.9 %	5.1 %	0.038	Asian
BRCA	ANO6	1.1 %	5.1 %	0.043	Asian
BRCA	PIK3CA	29.6 %	20.2 %	0.045	Black
BRCA	HAND2	0.9 %	5.1 %	0.046	Asian
BRCA	FBN2	1.1 %	5.1 %	0.047	Asian
BRCA	PPP1R3A	1.0 %	5.1 %	0.047	Asian
BRCA	ASH1L	2.3 %	8.5 %	0.049	Asian
BRCA	ROBO1	1.0 %	5.1 %	0.050	Asian

[a] Frequency (shortened as Freq) of mutation in White.

[b] Frequency of mutation in the minority group. Frequencies are calculated as the number of subjects who have at least one non-silent variant in this gene divided by the total number of subjects.

[c] Benjamini-Hochberg-adjusted p from logistic regression.

[d] The race of the minority group.

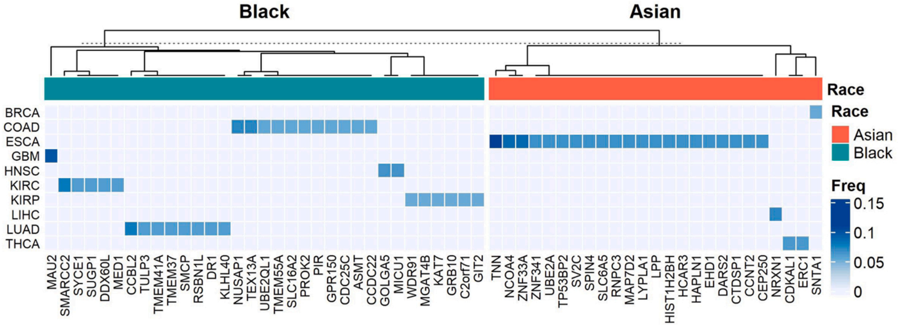

Fig. 1. Heatmap shows variant frequencies of 58 genes that had at least 5% variant frequencies for a cancer type in a minority ratial group. These genes had zero variant frequency in Whites.

greater than 10. We used logistic regression to investigate the relationship between heptatitis infection status and aflatoxin signature SBS24.

3. Results

3.1. Racial difference in single-gene variant frequency

We conducted an analysis of genomic variant frequencies across different racial groups, utilizing Whites as the reference cohort. This investigation identified a total of 63 significant genes (Table 1), encompassing several well-known oncogenic elements. Specifically, within the context of breast cancer, the TP53 oncogene exhibited a mutation frequency of 28.8% in Whites, 42.3 % in Blacks (adjusted $p = 0.005$), and 49.2 % in Asians (adjusted $p = 0.02$). Additionally, the transcription factor FOXO4, shown to both suppress and promote breast cancer progression [23] in breast cancer displayed a mutation frequency of 0.4% in Whites and 6.8 % in Asians (adjusted $p = 0.01$). Among the 63 significant genes, 61 were associated with breast cancer, while the remaining two were identified in kidney renal clear cell carcinoma. The prominence of breast cancer findings can be attributed to the increased statistical power derived from a larger sample size. A closer examination of the data revealed 58 gene-cancer instances where Whites show no mutation, and yet the minorities have a variant frequency > 5 % (Fig. 1).

3.2. Racial difference in mutational signature level

A mutational signature is the footprint of the carcinogenesis process. Deconvolving the mutation catalogs of TCGA patients against the 49 reference mutational signatures, we rendered a landscape of mutational signature intensities in around 10,000 cancer patients (Fig. 2A, Supplementary Table S2). Linear regression of mutational signature level on race identified 11 significant racial disparities with respect to separate mutational signatures, concerning diverse cancer types (Fig. 2B). Seven of the 11 results were with elevated mutational signatures in minority compared to Whites. The most significant result was observed for esophageal carcinoma (ESCA), where Asian patients have a significantly higher level of SBS16 compared to White patients (adjusted $p = 8.73 \times 10^{-19}$). The other racially disparate mutational signatures with a higher level in minorities include SBS17a, SBS24, SBS34, SBS16, SBS17a, and SBS34, most of which are of unknown etiology. SBS24 is a signature that signifies the exposure of aflatoxin, a well-known carcinogen for liver cancer. In our analysis, Asian liver cancer patients were found to receive higher contribution from SBS24 than Whites (adjusted $p = 0.002$), likely indicating a higher exposure to aflatoxin for Asian patients than Whites.

We also conducted logistic regression to detect whether a mutational signature is more represented in one race. The tests identified four significant associations between race and mutational signatures. In esophageal carcinoma, mutational signature SBS16 with unknown etiology is identified in 93.5 % of Asians compared to 41.2 % of Whites (adjusted $p = 9.85 \times 10^{-5}$). For the same mutational signature in head and neck squamous cell carcinoma, Blacks (78.7 %) have more presence than Whites (50.0 %) (adjusted $p = 0.016$); also in liver hepatocellular carcinoma, Asians (66.2 %) are significantly more than Whites (53.3 %) (adjusted $p = 0.018$). More interestingly, mutational signature SBS24, the signature of aflatoxin, a known carcinogen for liver cancer, is found in more Asians (82.8 %) than Whites (72.2 %) (adjusted $p = 0.018$).

Furthermore, logistic regressions were also conducted for a cancer type between distinct registry hospitals. This hospital-centered analysis revealed seven geospatial disparities for esophageal carcinoma, where signatures SBS7b, SBS16, and SBS17b were found to be significantly different between distinct hospitals (Table 2). To visually display the most drastic geospatial mutational signature difference in esophageal carcinoma patients, we plot the mutational signature proportion and race composition for esophageal carcinoma patients of three registry hospitals on a U.S. map: MD Anderson Cancer Center, ILSBio, and the University of Michigan (Fig. 2B). Signature SBS16 was observed in 94.74% of the subjects visiting ILSBio but was in only 29.17 % of the subjects visiting the University of Michigan. Such differences are primarily explained by disparate racial compositions of distinct medical facilities. Nearly 100 % of esophageal carcinoma patients of University of Michigan are White patients, while 100 % of ILSBio patients are from Asians.

3.3. Impacts of environmental pollutants on cancer survival and mutational signature level

In our investigation, we employed a multi-variate Cox survival model to scrutinize the prognostic implications of individual pollutants on patients' disease-specific survival. The analysis yielded 770 significant survival outcomes (Supplementary Table S3), with an enormously prevalent trend (96 %) indicating that a lower pollution level is associated with a more favorable prognosis. This observation underscores the robustness of our data, as it is less susceptible to statistical noise. Six noteworthy findings were selected for further analysis, including barium compounds in low-grade glioma, dimethylcarbamyl chloride, and hydrogen fluoride in colon adenocarcinoma, carbonyl sulfide and chlorimuron ethyl in liver hepatocellular carcinoma, and arsenic in skin cutaneous melanoma. Kaplan-Meier curves were constructed for these pollutants, as illustrated in Fig. 3, A-F. Many of the identified pollutants with significant prognostic impact are recognized carcinogens. For instance, arsenic, a heavy metal often present in contaminated water and food sources, has been established to elevate cancer risk, particularly in the context of skin cancer [24]. The observed significance of our results suggests that these pollutants may have influenced prognosis during treatment, or tumors developed under the influence of these pollutants may exhibit a more aggressive nature.

Next, we examined whether there are any associations between environmental pollutants and mutational signatures after adjusting for racial differences. Using stringent statistical criteria (details in Methods, Eqs. 1–4), our analysis identified 60 significant associations (adjusted

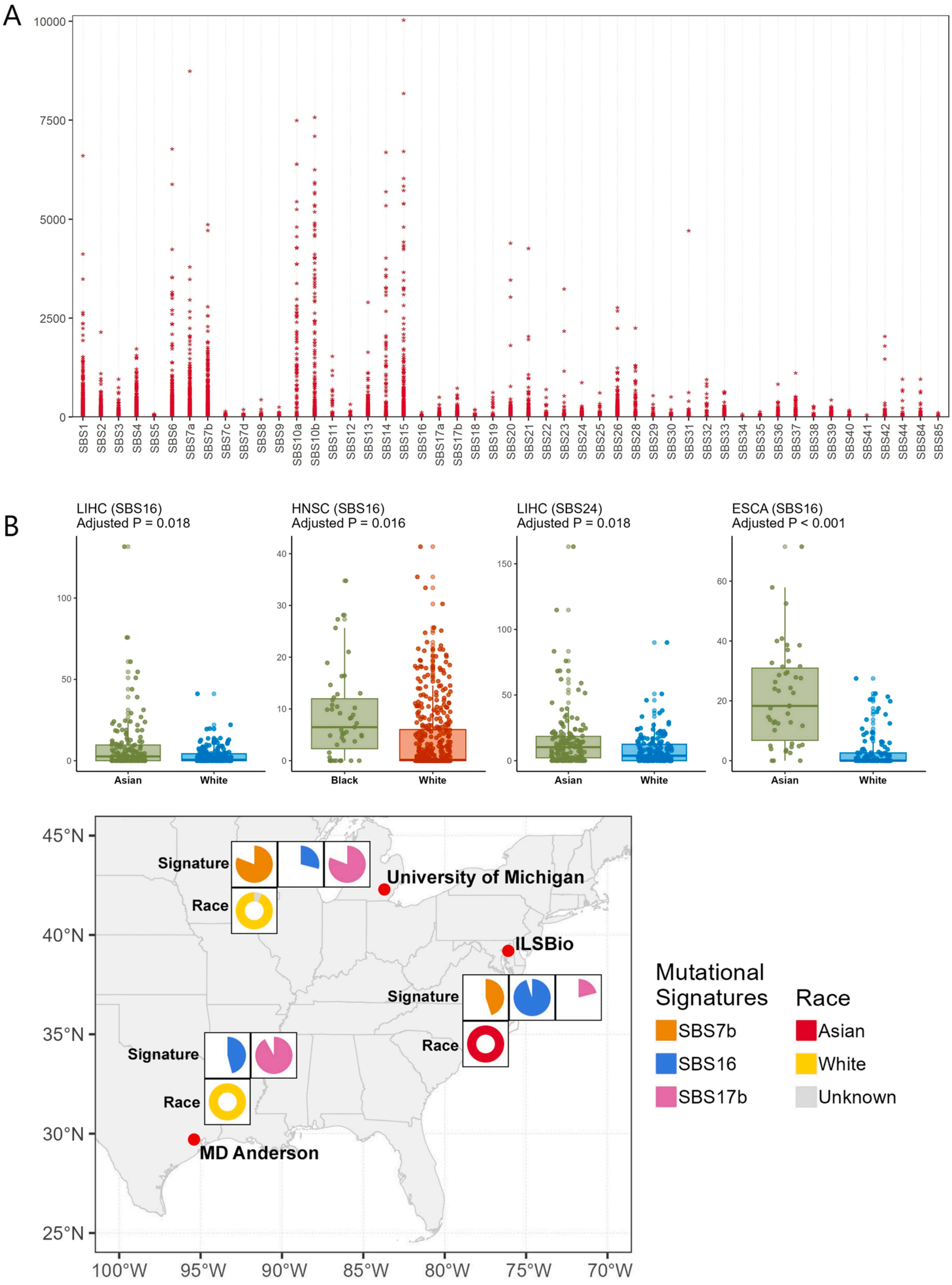

Fig. 2. Analysis results of mutational signatures and race. A. Overall mutational signatures of all patients. B. The four significant results where significant racial disparities were observed based on linear regression. Asians are denoted in green, Africans in red, and Whites (baseline) in blue. C. Esophageal carcinoma geographical etiology disparity among three cancer care units is entangled with racial disparity in esophageal carcinoma patients. Mutational signature compositions are displayed as pie charts, race compositions are displayed as donut charts.

$p < 0.01$), all of which occurred in association with stomach adenocarcinoma and SBS19 (Supplementary Table S4). All of the 60 significant associations affirm the direction that increased pollutant correlates with an increased level of a mutational signature. We selected six pollutants and plotted their association with mutational signature levels in barplots (Fig. 3, G-L). These results suggest environmental pollution with many known carcinogens may positively affect the tumorigenesis process. SBS19 has previously been associated with cobalt [25]. In our

Table 2
Significant results from logistic regressions that capture the difference of mutational signature between contributing hospitals for esophageal carcinoma patients.

Signature	Hospital1	Hospital2	Sample Size, Location 1	Sample Size, Location 2	Freq, Location 1[a]	Freq, Location 2[b]	Adjusted p[c]
SBS16	ILSBio	University of Michigan	38	48	94.74 %	29.17 %	1.04E-06
SBS17b	ILSBio	University of Michigan	38	48	21.05 %	81.25 %	1.26E-05
SBS17b	ILSBio	MD Anderson Cancer Center	38	13	21.05 %	92.31 %	2.41E-03
SBS17b	University of Michigan	Henry Ford Hospital	48	19	81.25 %	26.32 %	4.88E-03
SBS16	ILSBio	MD Anderson Cancer Center	38	13	94.74 %	46.15 %	2.25E-02
SBS17b	Henry Ford Hospital	MD Anderson Cancer Center	19	13	26.32 %	92.31 %	3.96E-02
SBS7b	ILSBio	University of Michigan	38	48	44.74 %	81.25 %	3.96E-02

[a] Frequency (shortened as Freq) of mutational signature in hospital 1.
[b] Frequency of mutational signature in hospital 2. Frequencies are calculated as the number of subjects who have a nonzero mutational signature value divided by the total number of subjects.
[c] Adjusted p from logistic regression, adjusted with the Benjamini-Hochberg procedure.

analysis, cobalt was statistically significant for several cancer and signature types, but did not meet our stringent threshold.

3.4. *Mutational Signature level in relation to oncogenic virus*

Infection with oncogenic viruses causes cancers. Cancer patients susceptible to an oncogenic virus may exhibit different mutational patterns than patients who are not infected. Of the 33 cancer types covered in TCGA, cervical squamous cell carcinoma and Head and Neck Squamous Cell Carcinoma are known to be influenced by Human Papillomavirus Virus (HPV), while liver hepatocellular carcinoma is known to be impacted by Hepatitis B or C Viruses (HBV or HCV).

We first examined the association between HPV status and mutational signature in CESC and HNSC. There is no significant difference in terms of HPV status among races: for HNSC cancer, 20 % of White patients, 10 % of Asian patients, and 11% of Black patients are HPV positive; for CESC, 95 % of White patients, 95 % of Asian patients, and 93 % of Black patients are HPV positive. From the CESC dataset and the HNSC cancer datasets, respectively, we identified 17 and 3 mutational signatures significantly associated with HPV status (Table 3). Interestingly, of the 20 total significant associations, 18 are negatively associated with HPV status, meaning HPV positive subjects have lower levels of certain mutational signature. The two positive associations involve SBS2 (AID/APOBEC family of cytidine deaminases) and SBS10b (polymerase epsilon exonuclease domain variants) in HNSC. These results suggest that HPV virus may induce cancer, but not directly through promoting excessive variants.

HBV and HCV are liver-specific viruses [26]. Using linear regression adjusted for race, we examined the association between mutational signatures and the infection status of HBV or HCV. The analysis found one significant result for SBS24: the signature of aflatoxin, with positive HCV correlating to a higher level of SBS24 (adjusted $p = 0.0003$). As in the cases of CESC and HNSC, the virus positive rates are not substantially different among races (HBV: Whites = 60 %, Blacks = 82 %, Asians = 63 %; HCV: Whites = 31 %, Blacks = 31 %, Asians = 25 %).

Since the aflatoxin signature SBS24 is related to a known liver cancer carcinogen, we performed additional in-depth analyses of SBS24 (Fig. 4A). The results show that Asian liver cancer patients have higher SBS24 levels than White patients ($p < 0.001$). Stratifying the race groups by virus infection status, Asian liver patients still show elevated SBS24 levels than Whites: Asian patients with HBV or HCV infection have higher SBS24 levels than White patients with the same infections (HBV $p = 0.008$, HCV $p = 0.035$), and Asian patients without hepatitis infection have higher SBS24 levels than White patients without hepatitis infection ($p = 0.024$). Comparing the SBS24 level between virus infection statuses, Asians and Whites show disparate phenomena: Asian patients with hepatitis infection (HBV or HCV) have higher SBS24 levels than Asian patients without hepatitis infection ($p = 0.028$), whereas no significant differences were observed between different hepatitis groups of White patients. These results suggest that virus-infected Asian liver cancer patients are more susceptible to aflatoxin exposure than vius-infected White patients. For Asian liver cancer patients, HBV or HCV infections exacerbate the aflatoxin mutational signature.

Furthermore, survival analysis was performed by incorporating hepatitis infection status and SBS24 in liver hepatocellular carcinoma, examining their interaction (Fig. 4B). HBV, HCV, and SBS24 exhibit associations with survival outcomes. Notably, infection with either HBV ($p = 0.031$) or HCV ($p < 0.001$) is correlated with a deteriorated prognosis, while lower SBS24 levels ($p < 0.005$) are associated with improved survival. However, the interaction term does not demonstrate a significant association with survival.

4. Discussion

Despite the longstanding efforts put to improve care for minority cancer patients, the evidence of cancer racial disparities is still strong. Instead of using traditional disparity measurements such as incidence rate and mortality, here we conducted a cancer desparity study linking environmental pollution to mutational signatures, intermediate surrogate phenotypes with traceable etiological mechanisms. Looking at mutational signatures is like looking at the history of tumorigenesis. Combining mutational signature data with environmental pollution data, we provide powerful evidence that pollutants can affect patients' overall mutational signature and prognosis. Through our analyses, we show that racial disparity has profound cascading effects. The disparate average socioeconomic statuses of different races often have a strong influence on the living conditions and locations, which subsumes the disparate types and degrees of pollutants inherent in the environments exposed to human subjects. First and foremost, we show that there are strong racial disparities in the variant frequencies of key oncogenes such as TP53. The variant frequency differences of these genes naturally translated into dispate mutational signature levels among different races, and indeed, we reported a plothera of Racial disparties in mutational signature levels regarding a multitude of cancer types. In particular, we pinpointed that the geospatial difference in SBS7b, SBS16, and SBS17b levels in esophageal carcinoma is overwhelmingly attributed to the drastically different racial composition of the patients visiting the different medical facilities.

The mutational signature SBS16, repeatedly appeared in our significant results. Currently (October 2023), the etiology is listed as unknown on the COSMIC website. However, SBS16 was previously associated with alcohol consumption in esophageal squamous cell carcinoma [27]. In our results, SBS16 was observed more in Asian than White for liver hepatocellular carcinoma and esophageal carcinoma. This is inconsistent with the previous report [27] that SBS16 is observed less in Asians. Also, a study has shown that alcohol consumption is least prevalent among Asian americans (38.0 %) compared to Whites (58.9 %) [28]. This suggests that SBS16 may represent multiple etiologies.

Performing cancer analyses at the mutational signature level enabled us to elucidate the disparate contribution of certain etiological factors to tumorigenesis in different races. For example, liver cancer has displayed an unambiguous racial difference, showing higher morbidity and

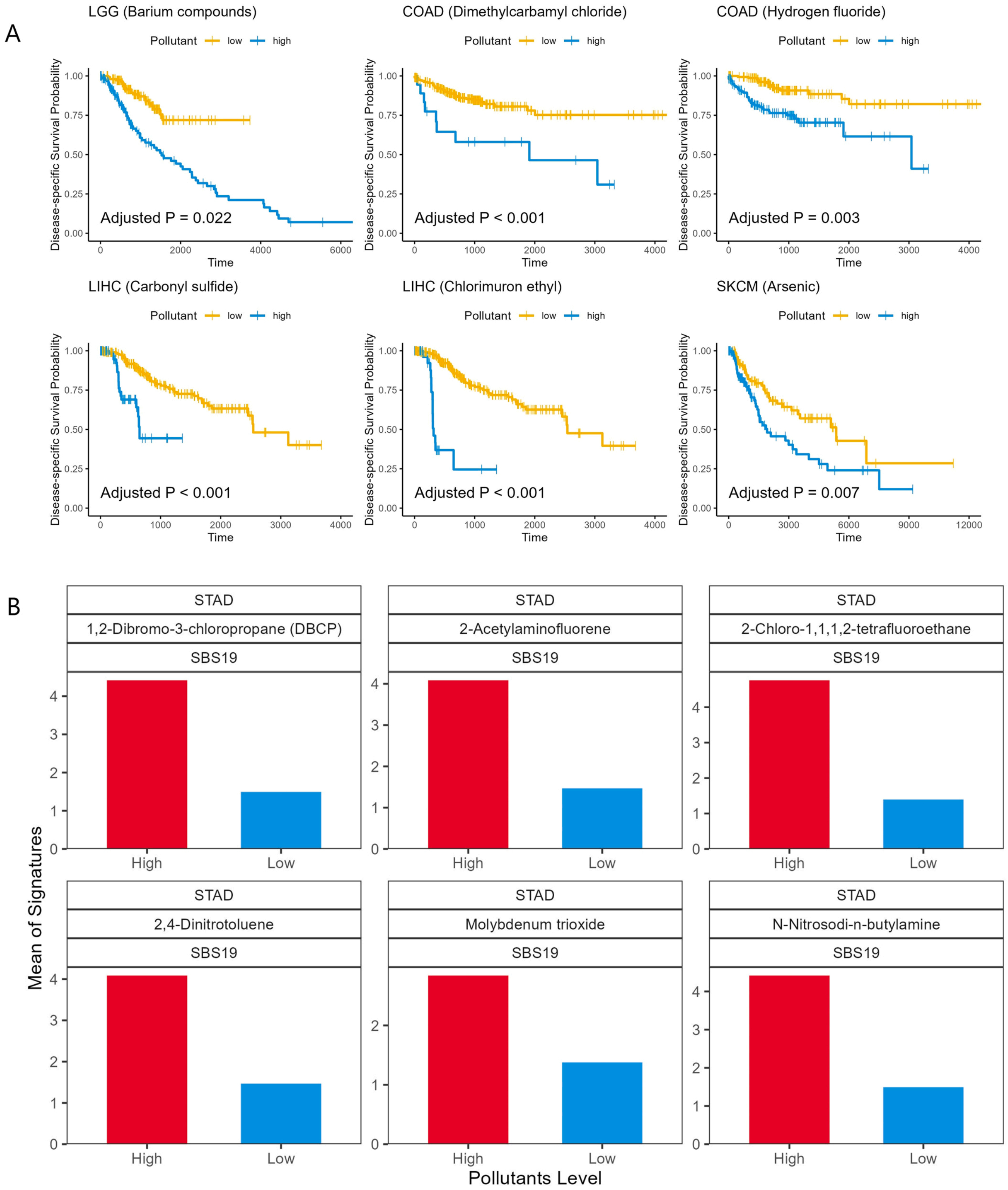

Fig. 3. Example pollutant analysis results. A. Six example survival plots that showing the increased pollution level is associated with poor prognosis. B. Six example bar plots that showing the increased pollution is associated with increased mutational signature level.

shorter survival in Asians than in Whites [29]. On the other hand, an array of risk factors have been associated with liver cancer, including virus infection of HBV and HCV, excessive alcohol intake, aflatoxin ingestion, and obesity. With the traditional approach that does not dissect the causes of liver cancer, no progress has been made to differentiate the etiology factors underlying ethnically different cancer patients. In our study, by quantifying the intensity of individual mutational signatures and modeling the intensity with race, we revealed that there is a racial difference in liver cancer in relation to aflatoxin. Aflatoxin is a carcinogen that commonly contaminates fermented foods and

Table 3
Mutational signatures that are significantly associated with HPV status in head and neck cancer and cervical cancer.

Signature	Cancer	Beta[a]	StdErr[a]	Adjusted p[a]
SBS2	HNSC	14.41	4.07	2.17E-02
SBS16	HNSC	-2.57	0.77	2.21E-02
SBS10b	HNSC	4.08	1.40	3.74E-02
SBS12	HNSC	-0.99	0.32	3.74E-02
SBS24	HNSC	-4.36	1.49	3.74E-02
SBS17a	HNSC	-0.74	0.26	4.08E-02
SBS10a	CESC	-97.94	16.94	5.39E-07
SBS3	CESC	-9.22	1.57	5.39E-07
SBS1	CESC	-213.97	39.90	1.82E-06
SBS10b	CESC	-162.19	29.88	1.82E-06
SBS17b	CESC	-10.86	2.03	1.82E-06
SBS21	CESC	-55.57	10.84	4.79E-06
SBS15	CESC	-433.08	90.35	1.95E-05
SBS26	CESC	-48.56	10.61	4.48E-05
SBS6	CESC	-260.99	58.42	6.45E-05
SBS14	CESC	-93.50	21.34	8.41E-05
SBS28	CESC	-42.16	10.22	2.04E-04
SBS33	CESC	-19.69	4.75	2.04E-04
SBS7b	CESC	-92.74	22.63	2.10E-04
SBS37	CESC	-13.61	3.34	2.14E-04
SBS44	CESC	-0.52	0.14	6.72E-04
SBS8	CESC	-3.41	1.02	2.83E-03
SBS12	CESC	-0.94	0.37	3.20E-02

bStardard error from linear regression.
cBenjamini-Hochberg-adjusted p from linear regression.
[a] Estimate/effect size from linear regression.

condiments, especially in hot and damp climates. Asians are more vulnerable to aflatoxin because fermented foods are more favorable among Asian [30].

Moreover, many developing countries in Asia are undergoing experience tropical climates, fostering the cultivation of crops that are particularly vulnerable to the proliferation of aflatoxins [31]. Chronic HBV infection and exposure to dietary aflatoxin play crucial roles in the multifaceted development of hepatocellular carcinogenesis, potentially acting synergistically. This synergy is manifested through the generation of DNA and protein adducts, accompanied by lipid peroxidation. Notably, individuals with hepatocellular carcinoma and HBV infection frequently demonstrate a prominent GC → TA transversion mutation at the third position of codon 249 in the p53 gene. Additionally, the HBx protein of HBV elicits diverse effects, including the promotion of cell cycle progression, augmentation of telomerase reverse transcriptase expression, deactivation of negative growth regulators, and suppression of the expression of p53 and other antiapoptotic tumor suppressor genes, along with factors associated with cellular senescence [32].

These disparate environmental exposure factors may account for the higher aflatoxin contribution to Asian liver cancer patients than White patients. Our analysis result of the higher susceptibility of aflatoxin in Asians can inspire race-specific preventative and intervention guidance. Beyond the elaborated liver cancer example, our analyses revealed a panel of elevated mutational signatures in minorities compared with the Whites. These pan-cancer analysis results can help combat cancer disparity in racial populations of diverse environmental exposure backgrounds.

Our study encountered two limitations. Firstly, the approximation of patient location based on the contributing hospital's location may

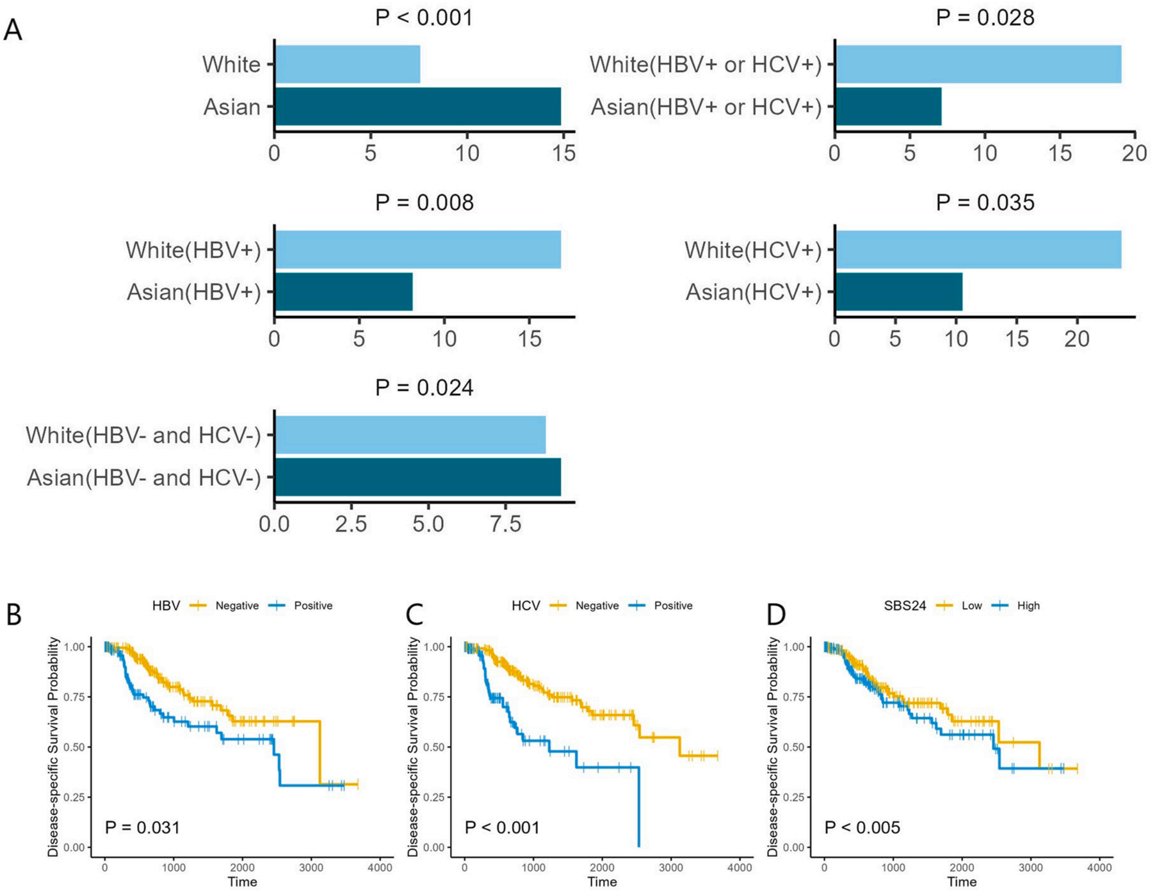

Fig. 4. A. Bar plots showing the results of SBS24 difference between Asian and White in liver hepatocellular carcinoma, grouping by different hepatitis infection status. B. Kaplan-Meier curve shows HBV infection is associated with worse survival. C. Kaplan-Meier curve shows HCV infection is associated with worse survival. D. Kaplan-Meier curve shows increased SBS24 level is associated with worse survival.

introduce inaccuracies, particularly for hospitals of national renown that attract patients from across regions. However, for the majority of patients, this approximation is deemed sufficiently accurate, and any associated noise is unlikely to impact statistical outcomes given a substantial sample size. Additionally, the environmental exposure data collected spanned from 2007 to 2017, preceding the initiation of the TCGA consortium around 2010. Most patients were recruited during the early stages of the consortium or prior to its inception. While it would be ideal to measure environmental exposures years before cancer development for optimal representation of mutational signature impact, the average exposure levels from 2007–2017 still effectively characterize the overall exposure during the preceding decade. To mitigate these limitations, rigorous statistical procedures were employed. Notably, our results exhibit remarkable consistency and alignment with conventional expectations. Specifically, 100% of significant outcomes indicate a negative association between pollutant levels and prognosis, and 96% demonstrate a positive correlation between pollutant levels and mutational signature levels. The evident coherence in our results attests to their reliability, particularly in the absence of undue noise within the datasets.

In addition to the results related to mutational signatures, our analyses also revealed several interesting findings. For large consortiums, data tracking is of critical importance. Most large consortiums lack detailed information, such as patients' residential locations, which forced us to interpolate the approximate location. Our analysis shows that race is strongly associated with contributing hospitals. For example, a hospital's contributing samples are of one single race. Such information is often neglected in pan-cancer analyses. Moreover, given the effort to represent minorities in consortium studies, some specific groups are still severely underrepresented. For example, Native Americans are virtually nonexistent in the TCGA or any other large consortiums. Hispanics are also very poorly represented in some of the TCGA's cancer types. This calls for more inclusive studies so the genomic characteristics of these underrepresented minorities can be closely examined to pinpoint the source of the disparity, thus allowing the implementation of more effective preventive measures.

5. Conclusion

There are four important findings from our study: 1) racial cancer difference can be observed at the mutational signature level; 2) mutational signatures are exacerbated by pollution exposure; 3) pollution exposures negatively affect patient survival; 4) Asian liver cancer patients were exposed to higher level of aflatoxin than Whites, and hepatitis infections are usually associated with a higher level of aflatoxin signatures. Our study is unique because it combines data from two governmental agencies and illustrates cancer racial disparity in the context of mutational signatures. Such a difference is just one of the many outcomes of the deep-rooted racial disparity problem in the U.S. Eradicating cancer racial disparities requires sustained efforts from government, and private organizations, as well as in-depth research that can shed light on racial genetics and genomics.

CRediT authorship contribution statement

All authors contributed to the manuscript. Guo designed and supervised the project. Guo secured funding and designed the study. Bai, Ma, Xia, Geng, Shen, Jiang and Yu conducted statistical analyses. Gong processed the EPA environmental data. Leng, Guo and Bai wrote the manuscript.

Declaration of Competing Interest

The authors declare no potential conflicts of interest.

Acknowledgment

We appreciated technical support by Mr. Jiapeng He from University of New Mexico. This study was supported by The Analytical & Translational Genomics Shared Resource and Bioinformatics Shared Resource of the University of New Mexico Comprehensive Cancer Center. This study was supported by Cancer Center Support Grant P30CA118100, P30CA240139, and R01ES030993–01A1 from the National Cancer Institute, USA. This study was supported by R01ES032037 from the National Institute of Environmental Health Sciences, USA.

Impact

This study demonstrates the links between mutational signatures and carcinogen exposures, including chemical pollutants and oncovirus infections.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.10.041.

References

[1] Pineros M, Znaor A, Mery L, Bray F. A global cancer surveillance framework within noncommunicable disease surveillance: making the case for population-based cancer registries. Epidemiol Rev 2017;39(1):161–9.
[2] Polite BN, et al. Charting the future of cancer health disparities research: a position statement from the American Association for cancer research, the American cancer society, the American society of clinical oncology, and the National Cancer Institute. Cancer Res 2017;77(17):4548–55.
[3] Ellis L, et al. Racial and ethnic disparities in cancer survival: the contribution of tumor, sociodemographic, institutional, and neighborhood characteristics. J Clin Oncol 2018;36(1):25–33.
[4] Butler SS, et al. Racial disparities in patient-reported measures of physician cultural competency among cancer survivors in the United States. JAMA Oncol 2020;6(1):152–4.
[5] Zwickl K, Ash M, Boyce JK. Regional variation in environmental inequality: industrial air toxics exposure in US cities. Ecol Econ 2014;107:494–509.
[6] Alvarez CH, Calasanti A, Evans CR, Ard K. Intersectional inequalities in industrial air toxics exposure in the United States. Health Place 2022;77.
[7] Liu J, et al. Disparities in air pollution exposure in the United States by race/ethnicity and income, 1990-2010. Environ Health Perspect 2021;129(12):1–14.
[8] Mikati I, et al. Disparities in distribution of particulate matter emission sources by race and poverty status. Am J Public Health 2018;108(4):480–5.
[9] Abel TD, Salazar DJ, Robert P. States of environmental justice: redistributive politics across the United States, 1993–2004. Rev Policy Res 2015;32:200–25.
[10] Bell M, Ebisu K. Environmental inequality in exposures to airborne particulate matter components in the United States. Environ Health Perspect 2012;120(12):1699–704.
[11] Kravitz-Wirtz N, Crowder K, Hajat A, Sass V. The long-term dynamics of racial/ethnic inequality in neighborhood air pollution exposure, 1990-2009. Du Bois Rev: Soc Sci Res Race 2016;13(2):237–59.
[12] Koh G, et al. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer 2021;21(10):619–37.
[13] Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature 2013;500(7463):415–21.
[14] Alexandrov LB, et al. The repertoire of mutational signatures in human cancer. Nature 2020;578(7793):94–101.
[15] Kucab JE, et al. A compendium of mutational signatures of environmental agents. Cell 2019;177(4):821–836 e16.
[16] Schrock AB, et al. Tumor mutational burden is predictive of response to immune checkpoint inhibitors in MSI-high metastatic colorectal cancer. Ann Oncol 2019;30(7):1096–103.
[17] Grimaldi D, Claessens YE, Mira JP, Chiche JD. Beyond clinical phenotype: the biologic integratome. Crit Care Med 2009;37(1 Suppl):S38–49.
[18] Russnes HG, et al. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. Sci Transl Med 2010;2(38):38ra47.
[19] Roth SH, Levanon EY, Eisenberg E. Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity. Nat Methods 2019;16(11):1131–8.
[20] Liu JF, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 2018;173(2):400.
[21] Jiang L, Yu H, Guo Y. Modeling the relationship between gene expression and mutational signature. Quant Biol 2023.
[22] Blokzijl F, Janssen R, van Boxtel R, Cuppen E. Mutational patterns: comprehensive genome-wide analysis of mutational processes. Genome Med 2018;10(1):33.
[23] Hornsveld M, et al. FOXO transcription factors both suppress and support breast cancer progression. Cancer Res 2018;78(9):2356–69.

[24] Mayer JE, Goldman RH. Arsenic and skin cancer in the USA: the current evidence regarding arsenic-contaminated drinking water. Int J Dermatol 2016;55(11): E585–91.

[25] Riva L, et al. The mutational signature profile of known and suspected human carcinogens in mice. Nat Genet 2020;52(11):1189–97.

[26] Ringehan M, McKeating JA, Protzer U. Viral hepatitis and liver cancer. Philos Trans R Soc B-Biol Sci 2017;372(1732).

[27] Moody S, et al. Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. Nat Genet 2021;53(11):1553 (-+).

[28] Delker E, Brown Q, Hasin DS. Alcohol consumption in demographic subpopulations: an epidemiologic overview. Alcohol Res 2016;38(1):7–15.

[29] McGlynn KA, Petrick JL, London WT. Global epidemiology of hepatocellular carcinoma: an emphasis on demographic and regional variability. Clin Liver Dis 2015;19(2):223–38.

[30] Sivamaruthi BS, Kesika P, Chaiyasut C. Toxins in fermented foods: prevalence and preventions – a mini review. Toxins 2018;11(1).

[31] Umar A, Bhatti HS, Honey SF. A call for aflatoxin control in Asia. Cabi Agric Biosci 2023;4:1.

[32] Moudgil V, Redhu D, Dhanda S, Singh J. A review of molecular mechanisms in the development of hepatocellular carcinoma by aflatoxin and hepatitis B and C viruses. J Environ Pathol Toxicol Oncol 2013;32(2):165–75.