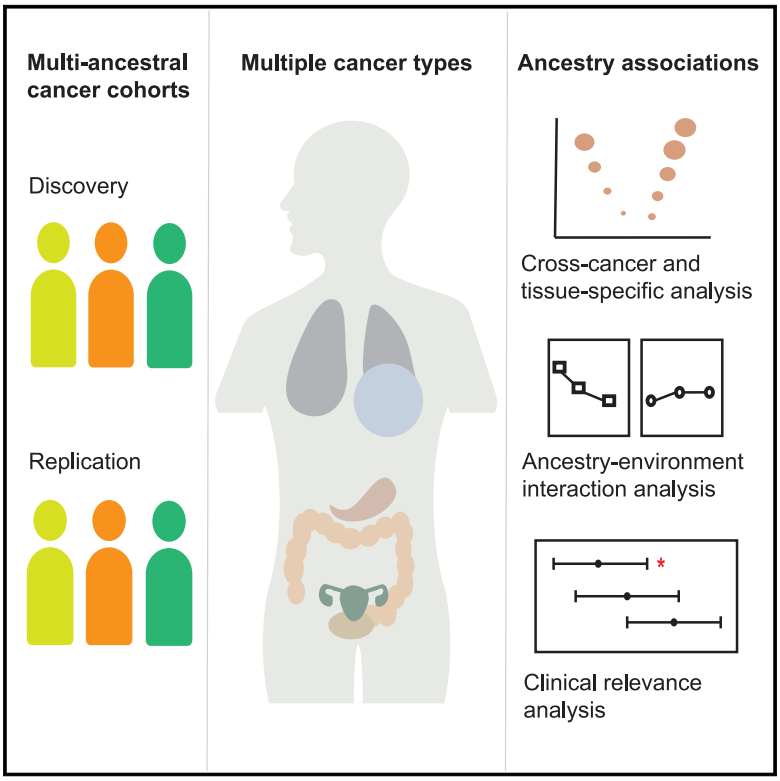
Cancer Cell

Article

Tumor sequencing of African ancestry reveals differences in clinically relevant alterations across common cancers

Graphical abstract



Highlights

- Describe African ancestry-associated cancer driver alterations in two large cohorts
- Cross-cancer association with African ancestry observed in *MYC* and *BRAF* genes
- Rates of actionable biomarkers vary in patients with African ancestry
- Smoking status interacts with *TP53* mutation rates in an ancestry-specific manner

Authors

Evelyn Jiagge, Dexter X. Jin, Justin Y. Newberg, ..., Garrett M. Frampton, Ethan S. Sokol, Jian Carrot-Zhang

Correspondence

ejiagge1@hfhs.org (E.J.), djin@foundationmedicine.com (D.X.J.), carrotj@mskcc.org (J.C.-Z.)

In brief

By leveraging two large genomic cohorts, Jiagge et al. characterize genomic alterations and clinically relevant biomarkers in patients with African ancestry across six common cancers. They identify an ancestry-environment interaction associated with driver alteration and highlight the need to increase representation of patients with African ancestry in clinical and research sequencing efforts.

Jiagge et al., 2023, Cancer Cell *41*, 1963–1971
November 13, 2023
https://doi.org/10.1016/j.ccell.2023.10.003

CellPress

Cancer Cell

CellPress
OPEN ACCESS

Article

Tumor sequencing of African ancestry reveals differences in clinically relevant alterations across common cancers

Evelyn Jiagge,[1,15,*] Dexter X. Jin,[2,15,*] Justin Y. Newberg,[2,15] Tomin Perea-Chamblee,[3,15] Kelly R. Pekala,[3,4,15] Christopher Fong,[3] Michele Waters,[3] David Ma,[3] Yvonne Dei-Adomakoh,[5] Gilles Erb,[6] Kanika S. Arora,[7] Sophia L. Maund,[8] Njoki Njiraini,[9] Atara Ntekim,[10] Susie Kim,[3] Xuechun Bai,[3] Marlene Thomas,[6] Ronwyn van Eeden,[11] Priti Hegde,[2] Justin Jee,[3,12] Debyani Chakravarty,[7,13] Nikolaus Schultz,[3] Michael F. Berger,[7,13] Garrett M. Frampton,[2] Ethan S. Sokol,[2] and Jian Carrot-Zhang[3,14,16,*]

[1]Hematology/Oncology Division, Department of Medicine, Henry Ford Health System, Detroit, MI, USA
[2]Computational Discovery, Foundation Medicine, Inc., Cambridge, MA, USA
[3]Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[4]Department of Surgery, Urology Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[5]Department of Haematology, University of Ghana Medical School, Accra, Ghana
[6]Global Product Development Medical Affairs – Oncology, F. Hoffmann-La Roche Ltd, Basel, Switzerland
[7]Marie-Josée and Henry R. Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer, New York, NY, USA
[8]Computational Sciences, Genentech, Inc., South San Francisco, CA, USA
[9]Department of Oncology, Kenyatta University Teaching Research and Referral Hospital, Nairobi, Kenya
[10]Department of Radiation Oncology, University of Ibadan, Ibadan, Nigeria
[11]Department of Medical Oncology, Chris Hani Academic Baragwanath Hospital, Johannesburg, South Africa
[12]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[13]Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[14]Clinial Genetics, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[15]These authors contributed equally
[16]Lead contact
*Correspondence: ejiagge1@hfhs.org (E.J.), djin@foundationmedicine.com (D.X.J.), carrotj@mskcc.org (J.C.-Z.)
https://doi.org/10.1016/j.ccell.2023.10.003

SUMMARY

Cancer genomes from patients with African (AFR) ancestry have been poorly studied in clinical research. We leverage two large genomic cohorts to investigate the relationship between genomic alterations and AFR ancestry in six common cancers. Cross-cancer type associations, such as an enrichment of *MYC* amplification with AFR ancestry in lung, breast, and prostate cancers, and depletion of *BRAF* alterations are observed in colorectal and pancreatic cancers. There are differences in actionable alterations, such as depletion of *KRAS* G12C and *EGFR* L858R, and enrichment of *ROS1* fusion with AFR ancestry in lung cancers. Interestingly, in lung cancer, *KRAS* mutations are less common in both smokers and non-smokers with AFR ancestry, whereas the association of *TP53* mutations with AFR ancestry is only seen in smokers, suggesting an ancestry-environment interaction that modifies driver rates. Our study highlights the need to increase representation of patients with AFR ancestry in drug development and biomarker discovery.

INTRODUCTION

Although African American (AA) patients with cancer have been underrepresented in most genomic studies, comparisons of African (AFR) ancestry and European (EUR) ancestry, based on limited data, suggest that the landscape of cancer driver alterations are different between populations.[1,2] For example, *MYC* amplification occurs more frequently in patients with prostate cancer of AFR ancestry.[3] In lung squamous cell carcinoma, there is increased genomic instability and homologous recombination deficiency in patients of AFR ancestry compared to patients of EUR ancestry.[4] In breast cancer, differences in outcomes between ancestry groups tend to be driven by the higher risk of developing biologically aggressive basal-like tumors or triple-negative breast cancer (TNBC) in patients of AFR ancestry compared to patients of EUR ancestry.[5,6] Although it remains unclear whether genomic differences are mediated by germline variations with different frequencies across different ancestries, or environmental exposures/social determinants related to race and ethnicity, the identification of ancestry-associated alterations underscores the need to better understand the underlying genomics of cancers from diverse populations.

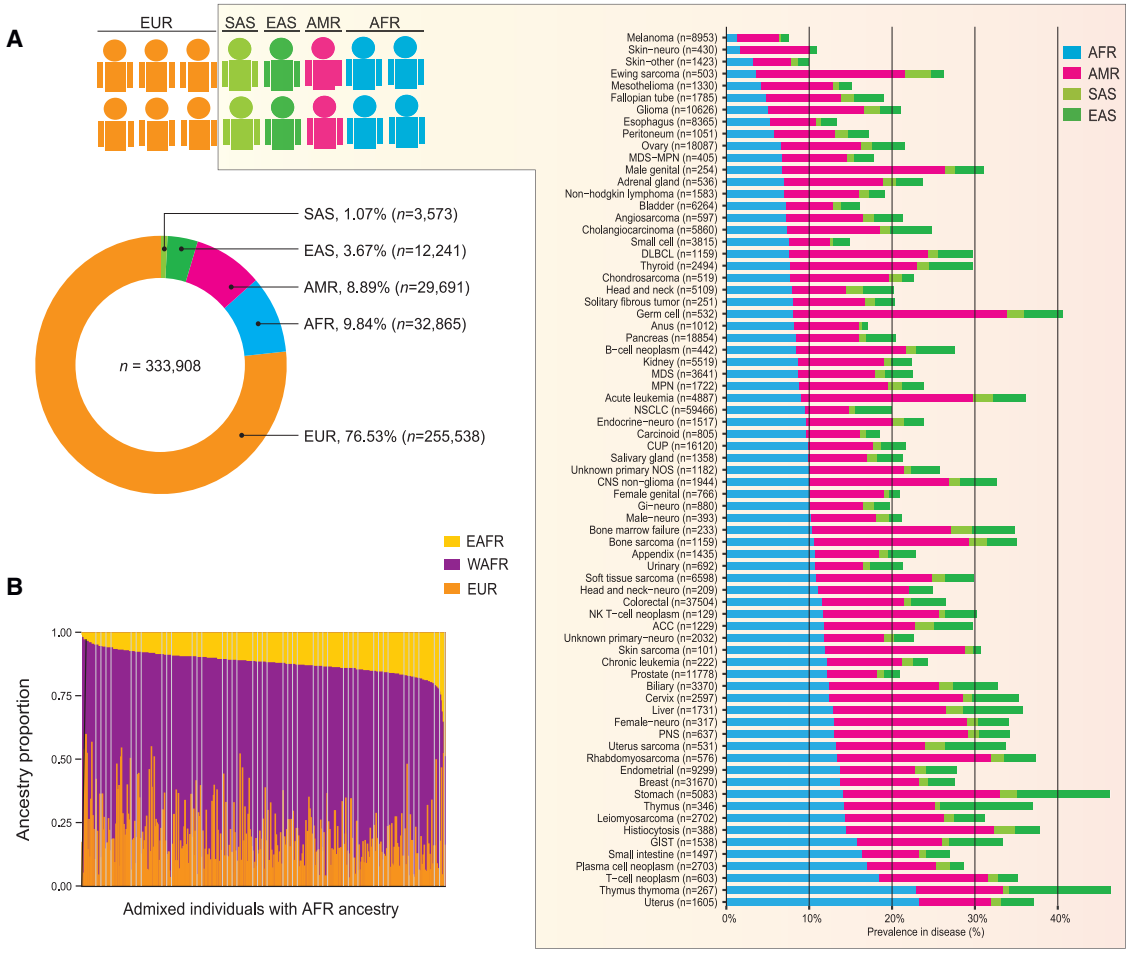


Figure 1. Ancestry representation across cancer types in the discovery cohort

(A) Overall patient population distribution across the ancestry cohorts–European (EUR), South Asian (SAS), East Asian (EAS), American (AMR), African (AFR) (left). Prevalence of ancestry cohorts within the diseases and sample size of each disease (right). Tumor types that are abbreviated are listed here: Adenoid Cystic Carcinoma (ACC), Central nervous system (CNS), Diffuse Large B Cell Lymphoma (DLBCL), Gastrointestinal (GI), Gastrointestinal Stromal Tumor (GIST), Myelodysplastic syndrome (MDS), Myelodysplastic-Myeloproliferative Neoplasm (MDS-MPN), Myeloproliferative Neoplasm (MPN), Natural Killer (NK), Not Otherwise Specified (NOS), Non-Small Cell Lung Carcinoma (NSCLC), Peripheral Nervous System (PNS).

(B) Fraction of East African (EAFR) ancestry, West African (WAFR) ancestry, and European (EUR) ancestry in each sample classified as Admixed AFR.

Cancer gene panel sequencing of tumors has become a common diagnostic tool to detect actionable genomic alterations, to guide treatment decisions, and to pair patients with appropriate clinical trials for novel molecular targeted therapy.[7–9] However, self-reported race and ethnicity information is not always available in genomic sequencing cohorts, limiting the use of real-world data for cancer disparity research. Moreover, race and ethnicity are categoric, social constructs, whereas genetic ancestry can be measured as a continuous variable, which may better represent the admixed nature of the AA population.[10] Accurate characterization and contextualization of ancestry-associated genomic alterations not only help to improve genomic diagnostic testing, but also provide insight into the development of targeted therapies, biomarkers, and personalized cancer care for patients from diverse populations.

Here, we infer genetic ancestry using cancer gene panel sequencing data for 333,908 tumor samples collected during routine clinical care in the US. We analyze driver alterations and biomarkers associated with AFR ancestry in six common cancers, adjusting for clinical and demographic covariates. We then use an independent sample cohort of 64,173 samples with paired tumor-normal sequencing data for validation. We address deficiencies in the understanding of clinically relevant alterations specific to patients with AFR ancestry who have access to diagnostic testing, and we demonstrate that caution should be taken when using prognostic biomarkers developed predominantly using patients with EUR ancestry.

RESULTS

Ancestry inference enabled analysis of African ancestry in large genomic sequencing cohorts

We used two datasets – 333,908 tumor-only samples served as our discovery cohort and 64,173 paired tumor-normal samples from Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) as our validation

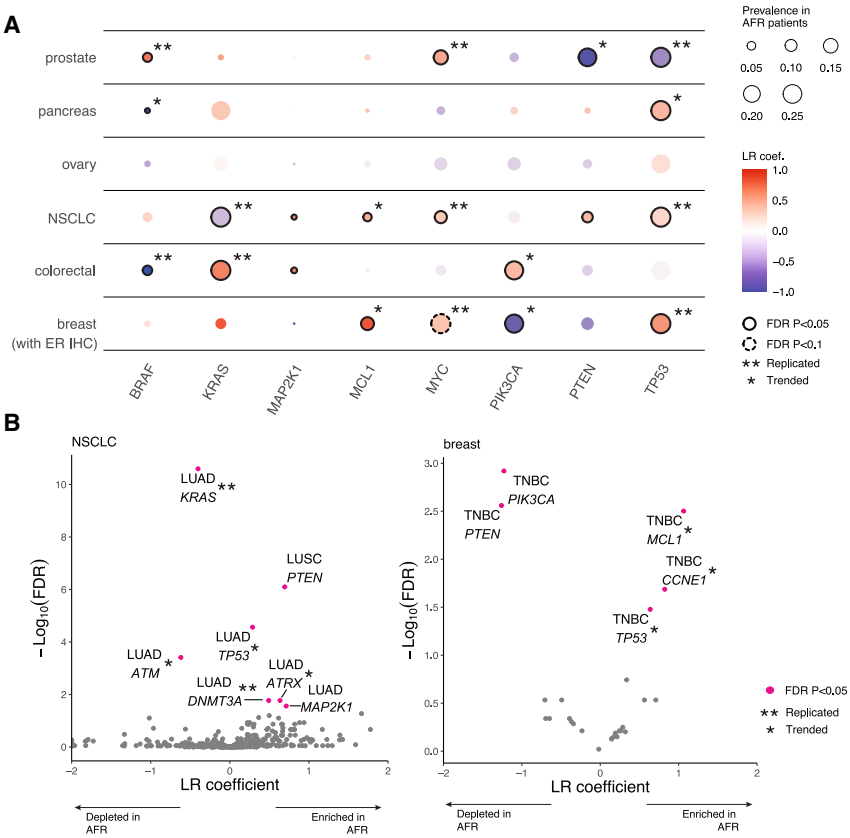


Figure 2. Cancer genes associated with African ancestry

(A) Distribution of gene alterations in each ancestry group across diseases. The size of each dot represents the prevalence of a given gene alteration in the African (AFR) cohort in the disease. Colors reflect the logistic regression coefficient (LR coefficient), where values >0 indicate a positive association with the African (AFR) ancestry percentage (red) and values <0 indicate a negative association with the AFR ancestry percentage (blue). The fully outlined circles passed the FDR correction of $p < 0.05$ and the partially outlined circles passed an FDR correction of $p < 0.1$.

(B) Left: Volcano plot of the association between percentage of AFR ancestry and gene alterations that were altered in at least 100 samples within the non-small cell lung cancer (NSCLC) – lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) subtypes. Right: Volcano plot of the association between percentage of AFR ancestry and gene alterations that were altered in at least 100 samples within the triple-negative breast cancer (TNBC) and ER + HER2- breast cancer subtypes. Replication with AFR ancestry in the MSK-IMPACT cohort is indicated by two asterisks ** (same LR coefficient direction and $p < 0.05$) and trended is indicated by one asterisk * (same LR coefficient direction).

cohort. In our discovery cohort, we inferred genetic ancestry based on 39,540 deeply sequenced germline SNP markers (STAR Methods), as described in prior studies.[2,3,11] Patients were classified into one of the five continental ancestry groups: AFR, EUR, American (AMR), South Asian (SAS), or East Asian (EAS). We subsequently inferred the percentage of West AFR ancestry (WAFR), East AFR ancestry (EAFR), and EUR ancestry for each individual across the discovery cohort (STAR Methods). We performed a ten-fold cross-validation on the 1000 Genomes samples (STAR Methods) and showed a high class-weighted precision (>99%) and recall (>99%) of our ancestry classifier. Moreover, an analysis of 1,184 paired AFR samples showed a high correlation of the WAFR ancestry in each pair (Pearson's $r = 0.97$, Figure S1A). These results suggest that continental and subcontinental ancestry can be accurately inferred from targeted gene panel sequencing of tumor-only DNA. In our replication cohort,[12] we also included clinical (stage, outcome), exposure (smoking history), and socioeconomic data (inferred income, average BMI, insurance status) linked to each sample.

We identified AFR ancestry in 9.8% (n = 32,865) of all samples in the discovery cohort (Figure 1A). Other ancestries in the dataset included: EUR: 76.5%, AMR: 8.9%, EAS: 3.7%, and SAS: 1.1% (Figure 1A). Moreover, the AFR cohort displayed an admixture of predominately the WAFR and EUR ancestry (Figure 1B; Figure S1B), consistent with the underlying genetic heterogeneity in the AFR samples. The sequenced cohort had a distribution across cancer types that mirrored that of population epidemiology of the United States,[13,14] such as the enrichment of prostate, stomach, and endometrial cancer, and the depletion of ovarian and esophagus cancer with AFR ancestry (Figure S2 and Table S1). However, discrepancies between our study cohort and population-based cancer incidence data are expected, as certain cancer types/subtypes might be guided more toward testing in search of actionable mutations. Taken together, we identified a large cohort of patients of cancer with AFR ancestry (discovery cohort) who received clinical genomic testing that would be otherwise ignored due to the lack of reported race or ethnicity. Cancer types with at least 1,000 samples of AFR ancestry, including non-small cell lung cancer (NSCLC), colorectal cancer, breast cancer, prostate cancer, pancreatic cancer, and ovarian cancers, were the focus of this study since they provided sufficient power to detect differences of low-frequency alterations in AFR ancestry.

Gene-level analysis revealed distinct drivers in patients with AFR ancestry

Given the nature of admixture in our AFR cohort, which was predominantly WAFR[15] (Figure 1B), we termed this as the AFR ancestry percentage in downstream correlative analyses. Additionally, since WAFR ancestry calls were not available for the validation cohort, AFR ancestry was used as a surrogate (STAR Methods). We associated the percentage of AFR ancestry per individual with 253 cancer genes (Methods S1), using a logistic regression model for the presence or absence of an oncogenic alteration in the gene, while accounting for age, sex, tumor mutational burden (TMB), subtype, and panel version (STAR Methods). Only known or likely functional genomic alterations in genes of interest were included in the gene-level analysis

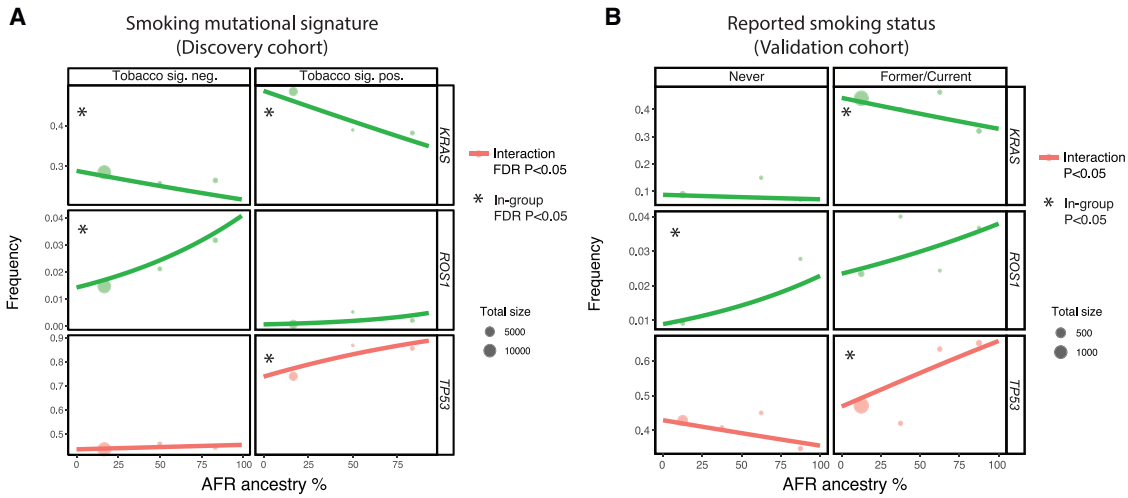


Figure 3. Ancestry-smoking interaction associated with driver mutations in lung cancer

(A) Logistic regression showing the associations between the percentage of African (AFR) ancestry and *KRAS*, *ROS1* fusion, and *TP53* in the discovery cohort in a smoking signature positive (right) and smoking signature negative (left; tumor mutational burden [TMB] < 5) cohort. Points represent the prevalence of alterations within AFR ancestry percentage tertile. One asterisk * indicates FDR $p < 0.05$ in the stratified analysis. Red lines indicate that the ancestry-smoking interaction term was statistically significant, FDR $p < 0.05$.

(B) Logistic regression showing the associations between the percentage of AFR ancestry and *KRAS*, *ROS1* fusion, and *TP53* in the validation cohort in a former/current smoker (right) and never smoker (left) cohort. Points represent the prevalence of alterations within AFR ancestry signature quartiles. One asterisk * indicates $p < 0.05$ in the stratified analysis. Red lines indicate that the ancestry-smoking interaction term was statistically significant, $p < 0.05$.

(STAR Methods). As the prevalence of multiple cancer subtypes was different between AFR ancestry and EUR ancestry, our analyses were performed across subtypes and within subtypes. Our initial analysis identified eight genes that were significant (false discovery rate [FDR] $p < 0.05$, Table S2) in at least two cancer types studied and used the MSK-IMPACT cohort for validation adjusting for age, sex, stage, subtype, and smoking status where applicable (Figure 2A). We replicated previously reported cross-cancer associations, such as *TP53* and *PIK3CA*,[1] and showed additional cross-cancer associations, such as the enrichment of *MYC* amplifications in NSCLC, prostate cancer, and breast cancer with AFR ancestry (Figures 2A, and S3A, and Table S3). Notably, *MYC* amplifications were associated with worse overall survival in those cancers (Figure S3B). Moreover, depletions of *BRAF* alterations (including V600E and fusions) were seen in colorectal and pancreatic cancers, whereas *KRAS* mutations (including G12C) were depleted in NSCLC and resistance mutations (including G12D, G13D, G12V, and Q61H) were enriched in colorectal cancer with AFR ancestry (Figures 2A, S4, and Table S4). In addition, cancer type and subtype-specific analysis of AFR ancestry revealed an enrichment of *TP53* in pancreatic cancer, a positive correlation of *DNMT3A* and a negative correlation with *KRAS* in lung adenocarcinoma (LUAD), and an enrichment of *MCL1* in TNBC (Figure 2B, Table S2A, and S2B). Taken together, ancestry analysis in multiple cancer types may allow for insightful application of existing treatment paradigms with targeted therapy across cancer types for patients with AFR ancestry.

Leveraging the large sample size in our discovery cohort, we investigated driver alterations in NSCLC, specifically LUAD, where multiple studies have shown inconsistent findings of their associations with AFR ancestry.[16–20] When separately analyzing patients with LUAD exposed vs. not exposed to smoking in both the discovery cohort and the validation cohort (STAR Methods), we found that *ROS1* fusions were enriched with AFR ancestry in patients negative for the smoking mutational signature and never smokers of AFR ancestry (Figures 3A and 3B). *KRAS* mutations were depleted with AFR ancestry, independent of smoking exposure (Figures 3A and 3B). Interestingly, we found that driver mutations in *TP53* were only enriched in ever smokers and patients positive for the smoking signature, but not in never smokers and patients without the smoking signature (Figures 3A and 3B). We subsequently tested the interaction of AFR ancestry and smoking exposure and its association with LUAD drivers (STAR Methods). We identified an interaction of AFR ancestry with the smoking signature (FDR $p = 0.0003$, interaction coefficient = 0.78) and an interaction of AFR ancestry with reported smoking status (FDR $p = 0.007$, interaction coefficient = 1.07) that were associated with *TP53* mutations. The interaction of AFR ancestry with smoking status remained significant after accounting for average BMI, inferred income, and insurance type ($p = 0.004$, interaction coefficient = 1), which was available in the MSK-IMPACT cohort. The ancestry-smoking interaction analysis leveraging two independent cohorts suggests that ancestry modifies the effect of smoking exposure on developing *TP53*-mutant LUAD tumors and may lead to different smoking-related pathogenesis in patients of AFR ancestry.

Clinically relevant biomarkers showed different prevalences in patients with AFR ancestry

Next, we investigated biomarkers used to guide therapeutic decisions in patients with AFR ancestry, such as TMB and microsatellite instability (MSI) status for immune checkpoint inhibitors, and homologous recombination repair (HRR) deficiency for poly-ADP ribose polymerase inhibitors,[21,22] adjusting for age, sex, subtype, and panel version. Of note, the TMB calculation we

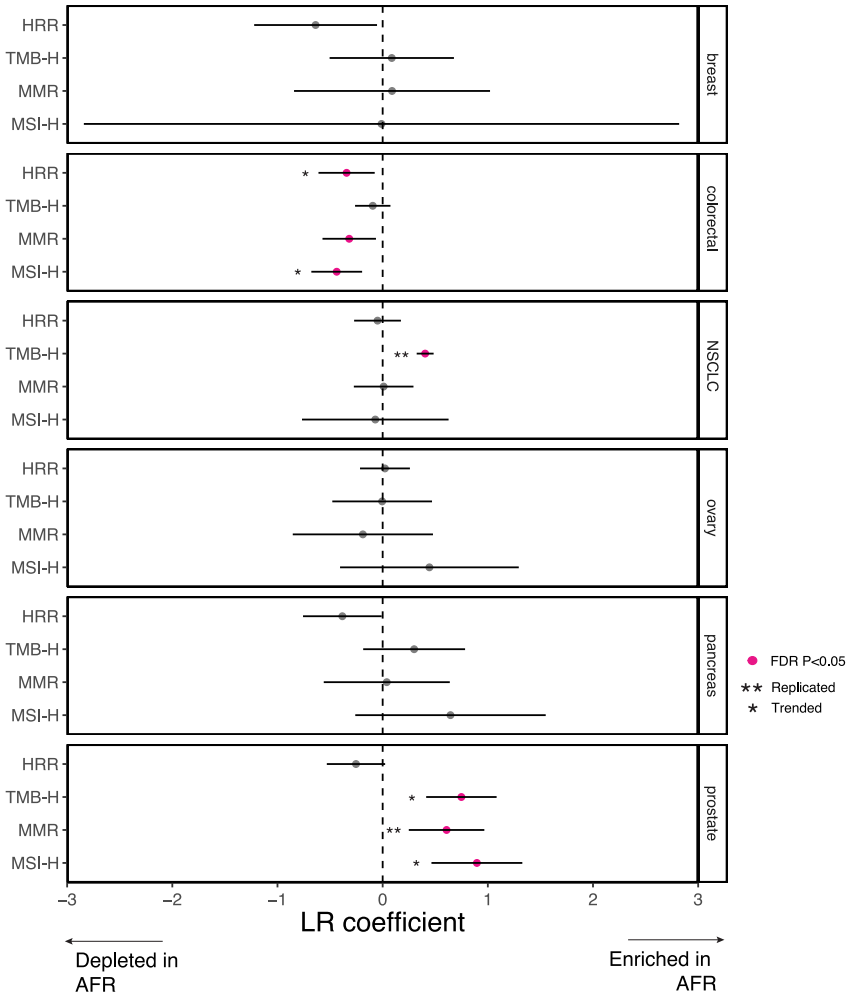


Figure 4. Biomarkers associated with African ancestry

Forest plots of the logistic regression coefficients and their confidence intervals showing the associations between African (AFR) ancestry percentage and four biomarkers (homologous recombination repair [HRR] gene set alterations [*BRCA1/2*], tumor mutational burden-high [TMB-H], mismatch repair [MMR] gene set [*MLH1/MSH2/MSH6/PMS2*], and microsatellite instability-high [MSI-H]) across breast cancers, colorectal cancers, non-small cell lung cancers (NSCLC), ovarian cancers, pancreatic cancers, and prostate cancers. Points colored in pink indicate significant values, FDR $p < 0.05$. Bars indicate the 95% confidence intervals. Replication with AFR ancestry in the MSK-IMPACT cohort is indicated by two asterisks ** (same logistic regression [LR] coefficient direction and $p < 0.05$) and trended is indicated by one asterisk * (same LR coefficient direction).

used excluded mutations that were predicted to be germline within an individual patient, as well as mutations that were recurrently predicted to be germline within each ancestry cohort (STAR Methods). These steps were used to minimize the TMB inflation for non-EUR samples where germline variants cannot be sufficiently filtered due to the lack of non-EUR samples in the reference databases.[23] Furthermore, the MSK-IMPACT cohort with matched normal sequencing was used for validation to ensure that the observed associations were not confounded by tumor-only sequencing.

TMB-high status (>=10 mutations/megabase) was more common in patients with AFR ancestry in NSCLC and prostate cancer (Figure 4 and Table S5). The increase of TMB-high cases with AFR ancestry remained significant in the MSK-IMPACT lung cohort of smokers (logistic regression $p = 0.023$, odds ratio [OR] = 1.6), but not in never smokers. HRR genes *BRCA1* and *BRCA2* were less frequently mutated with AFR ancestry in colorectal cancer (Figures 4, S5, and Table S5). MSI-high was observed at a lower frequency in colorectal cancer and at a higher frequency in prostate cancer with AFR ancestry (Figure 4 and Table S5). Similar with the trends observed for the MSI status, patients with AFR ancestry had a lower mutation rate in mismatch repair genes (MMR) including *MLH1*, *MSH2*, *MSH6*, and *PMS2* in colorectal cancer and a higher rate in prostate cancer (Figure 4 and Table S5).

We then analyzed the actionability of AFR ancestry-associated genomic alterations, which were annotated using OncoKB.[24] We observed U.S. Food and Drug Administration (FDA)-approved, actionable alterations enriched in patients with AFR ancestry including *ROS1* fusion in NSCLC negative for smoking signature, and *CDK12* alteration and *NTRK3* fusion in prostate cancer (Figure 5 and Table S6). There was also an enrichment of *KRAS* mutations (G12D, G13D, G12V, and Q61H) in colorectal cancer of AFR ancestry conferring resistance to standard-of-care treatments (Figure 5 and Table S6). *KRAS* G12C, with recently approved inhibitors, was depleted in both smoking and non-smoking cohorts. Interestingly, although most actionable mutations in *EGFR* were enriched with AFR ancestry, *EGFR* L858R, a key FDA-recognized biomarker in NSCLC predictive of response to all standard-of-care *EGFR* tyrosine kinase inhibitors, tended to be depleted with AFR ancestry (Figure 5 and Table S6).

Finally, we investigated whether there was an ancestry-specific association between survival and the genomic features associated with AFR ancestry. This analysis sought to identify a potential prognostic marker that might be generalizable across ancestries, or ancestry specific. For each cancer type, we tested the interaction of AFR ancestry and its associated genes on survival, using a Cox model followed by FDR correction (STAR Methods). The interaction indicates whether the influence of a gene on survival depends on ancestry. We did not detect any significant interactions, although there was a borderline association of overall survival with the interaction of *CCNE1* amplification and AFR ancestry in breast cancer (FDR $p = 0.11$, OR = 3.3, Figure 6 and Table S7), adjusting for age, disease stage, subtype, fraction of genome altered, inferred income, BMI, and

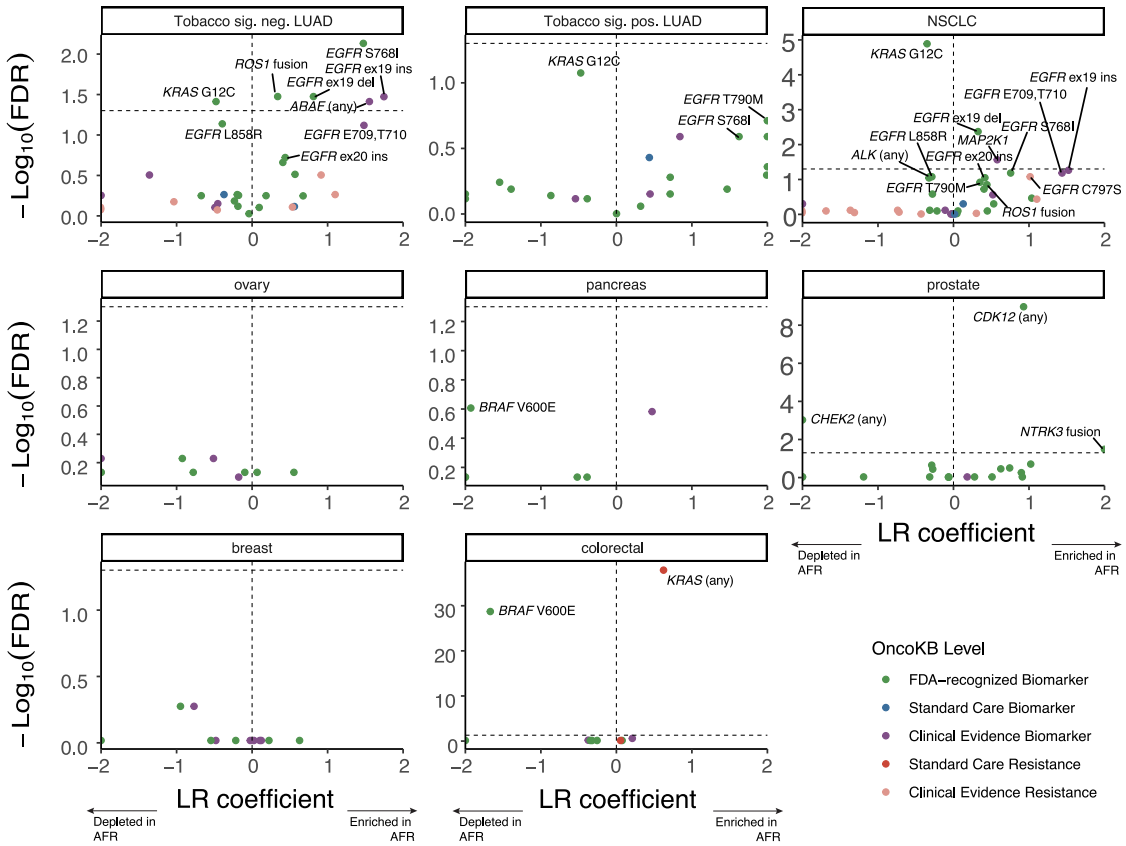


Figure 5. Clinically actionable alterations associated with African ancestry

Volcano plots of the association between African (AFR) ancestry percentage and alterations annotated with OncoKB actionability in smoking signature-negative and TMB <5 lung adenocarcinomas (LUAD), smoking signature-positive LUAD, all non-small cell lung cancers (NSCLC), ovarian cancer, pancreatic cancer, prostate cancer, breast cancer, and colorectal cancer.

insurance status (STAR Methods). Moreover, we specifically tested TMB in patients with LUAD who were treated with immunotherapy and confirmed TMB-high status as an outcome biomarker for immunotherapy in patients of AFR ancestry (Figure S6). Due to the limited sample size of patients with linked clinical outcome data, we cannot conclude that *CCNE1* amplification is an ancestry-specific prognostic marker, but we suggest further analysis of outcome prediction by a biomarker accounting for ancestry. Taken together, we demonstrated differences in clinically actionable mutations in multiple cancer types of AFR ancestry, highlighting the need to ensure equal access to diagnostic testing and precision oncology.

DISCUSSION

Genomic differences between patients from different populations have been analyzed using existing genomic datasets, based on self-reported race/ethnicity[19,25–27] or genetic ancestry inferred from sequencing data.[2,3,11] Here, we used genome-wide SNPs targeted in clinical tumor panel sequencing to infer genetic ancestry for cancer samples of which race information was not available[2,3,11] that enabled us to assemble a large cancer cohort of AFR ancestry. Moreover, considering ancestry as a continuous variable allowed us to include patients with mixed ancestries. We then used a clinical oncology cohort with matched normal sequencing for validation and to control for additional confounders such as subtype biases and metastatic status. We confirmed multiple ancestry-related differences in driver genes and identified several differences across cancer types, such as the depletion of *BRAF* mutations in colorectal and pancreatic cancers, and the enrichment of *MYC* amplification in lung, breast, and prostate cancers that may partially explain outcome disparities in patients with AFR ancestry.

We showed clinically actionable alterations enriched in patients of AFR ancestry that were not previously reported, such as *ROS1* fusion in non-smoking patients with LUAD and *NTRK3* fusion in prostate cancer. TMB-high cases were significantly enriched in NSCLC smokers and prostate cancer, which was further replicated in an independent cohort where TMB was calculated from paired tumor-normal sequencing. We confirmed that TMB-high status was associated with better immunotherapy outcome in patients of AFR ancestry, highlighting the importance of improving access to genomic testing and precision medicine for those patients. In contrast, we showed *CCNE1* amplification as a putative ancestry-specific prognostic indicator, highlighting the importance of increasing ancestral diversity in clinical trial design.

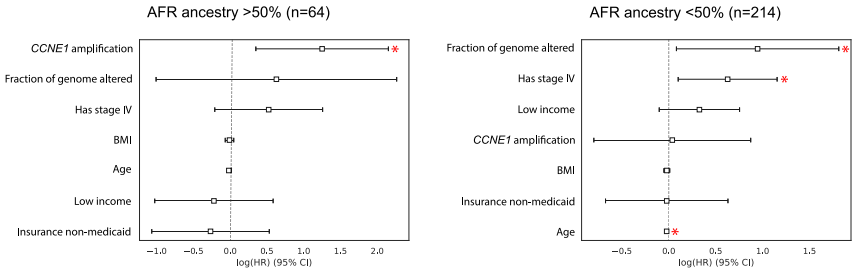


Figure 6. Ancestry specificity of *CCNE1* amplification as a biomarker

Visualization of the interaction of African (AFR) ancestry percentage and *CCNE1* amplification on survival in patients with triple-negative breast cancer (TNBC). The natural log of hazard ratio (HR) for overall survival with AFR ancestry >50% (left) and AFR ancestry <50% (right) is shown. Bars indicate the 95% confidence intervals. Asterisks represent significant associations ($p < 0.05$) with survival in the stratified analysis using the multivariate Cox regression model.

We further investigated the relationship between genetic ancestry and environmental exposure with a focus on LUAD. For patients with tobacco smoking-induced mutational processes, we found distinct genetic and potential therapeutic vulnerabilities in AFR ancestry compared to EUR ancestry—smokers with AFR ancestry were marked by *TP53* loss but were less likely to develop *KRAS* activation-driven lung cancers, despite lower smoking pack years compared to smokers with EUR ancestry in our cohort ($p = 0.02$, OR = 0.6).[28–30] On the other hand, never smokers with AFR ancestry were mainly driven by RAS signaling activation, similar to patients of EUR ancestry with a higher frequency of *ROS1* fusions compared to the EUR ancestry. Although we cannot conclude that the dependency of AFR ancestry on smoking-related *TP53* mutations is due to different smoking patterns or ancestry-specific genetic variations—perhaps a locus-specific ancestry (local ancestry) analysis is needed to understand the heritability of this interaction—our study revealed an ancestry-environmental interaction that can modify tumorigenesis. This finding emphasizes the importance of including diverse patient populations to better understand environmental-induced tumor biology, and the urgent need to develop new drug targets for patients with AFR ancestry and other non-EUR ancestry.

One limitation of our study is that all samples were collected from patients in the US. We cannot assess genomic differences in patients who reside in different geographic regions and are affected by different environmental factors, such as environmental and chemical pollution. Moreover, we cannot assess the alteration prevalence in patients who do not have access to genomic testing, and the survival data were collected only from patients treated in a National Cancer Institute-designated cancer center. These limitations exclude the majority of cancer patients with AFR ancestry. Another limitation is that our discovery cohort is a tumor-only sequencing cohort, where in some cases (e.g., the HRR genes and the MMR genes), pathogenic somatic and germline mutations were assessed together, although we can still understand the clinical utility of the ancestry-associated biomarkers, be it somatic or germline. Finally, our study focused on alterations in known cancer genes and pathogenic alterations that were mainly discovered in patients of non-AFR ancestry and are incorporated into clinical sequencing platforms. Therefore, efforts should be made to increase diversity in both clinical and research sequencing to close the gap of our knowledge about tumor genomes of patients with AFR ancestry.

In summary, our study uncovered ancestry-associated driver alterations and biomarkers across multiple cancer types and subtypes. We showed that it is important to jointly investigate genetic ancestry, environmental exposure, and ancestry-environment interaction in understanding genomic differences between different ancestral groups. We demonstrated the power of using real-world sequencing data that opens the door to large-scale interrogation of disparities in cancer genomics, and we call for the routine and expanded inclusion of patients with AFR ancestry in future genomic studies and clinical trials to improve diagnostics and precision medicine for all.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Ancestry assignment
  - Genomic profiling and alteration detection
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Identification of ancestry association
  - Clinical data analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2023.10.003.

ACKNOWLEDGMENTS

J.C.-Z. was funded by the National Cancer Institute 4R00CA259223, 1R21CA280577, and U54CA137788 & U54CA132378. K.R.P. was funded by the Ruth L. Kirschstein National Research Service Award (T32-CA082088) and National Institutes of Health/National Cancer Institute (P30-CA008748). Parts of this study were funded by F. Hoffmann-La Roche Ltd, Basel, Switzerland. Research support in the form of third-party editorial assistance, furnished by Katie Wilson, PhD, at Health Interactions was provided by F. Hoffmann-La Roche Ltd.

AUTHOR CONTRIBUTIONS

Conceptualization, E.J., D.X.J., J.Y.N., Y.D.-A., G.E., S.L.M., M.T., R.v.E., P.H., G.M.F., E.S.S., and J.C.-Z.; methodology, D.X.J., J.Y.N., T.P.-C., M.F.B., G.M.F., E.S.S., and J.C.-Z.; formal analysis and investigation, E.J., D.X.J., J.Y.N., T.P.-C., K.R.P., C.F., M.W., D.M., K.S.A., S.L.M., N.N., A.N.,

S.K., X.B., J.J., D.C., N.S., M.F.B., G.M.F., E.S.S., and J.C.-Z.; writing – original draft, D.X.J. and J.C.-Z.; writing – review and editing, E.J., D.X.J., J.Y.N., T.P.-C., K.R.P., C.F., M.W., D.M., Y.D.-A., G.E., K.S.A., S.L.M., N.N., A.N., S.K., X.B., M.T., R.v.E., P.H., J.J., D.C., N.S., M.F.B., G.M.F., E.S.S., and J.C.-Z.; supervision, D.X.J. and J.C.-Z.

DECLARATION OF INTERESTS

E.J. reports consulting fees for a Ghanaian breast cancer patient care pathway from Genentech, Inc. D.X.J., J.Y.N., G.M.F., and E.S.S. are employees of Foundation Medicine, Inc., a member of the Roche group, and hold stock in F. Hoffmann-La Roche Ltd. Y.D.-A. has received a grant from the National Heart, Lung, and Blood Institute in the USA, outside of the submitted work. G.E. and M.T. are employees of and hold stock in F. Hoffmann-La Roche Ltd. S.L.M. is an employee of Genentech, Inc., a member of the Roche group, and holds stock in F. Hoffmann-La Roche Ltd. N.N. has received honoraria for Breast Preceptorship training and the KESHO meeting, and is a member of a Global Advisory Board. A.N. has received honoraria for lectures and investigational products for clinical trials from F. Hoffmann-La Roche Ltd. R.v.E. has received honoraria for lectures and presentations from F. Hoffmann-La Roche Ltd, Takeda, MSD, Boehringer Ingelheim, and AstraZeneca, has received support for attending meetings and/or travel from F. Hoffmann-La Roche Ltd, Janssen, and MSD, and has participated in a data safety monitoring board or advisory board for Takeda and F. Hoffmann-La Roche Ltd. All authors received research funding in the form of third-party editorial support from F. Hoffmann-La Roche Ltd.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper received support from a program designed to increase minority representation in their field of research.

Received: October 25, 2022
Revised: August 2, 2023
Accepted: October 4, 2023
Published: November 13, 2023

REFERENCES

1. Yuan, J., Hu, Z., Mahal, B.A., Zhao, S.D., Kensler, K.H., Pi, J., Hu, X., Zhang, Y., Wang, Y., Jiang, J., et al. (2018). Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. Cancer Cell *34*, 549–560.e9.
2. Carrot-Zhang, J., Chambwe, N., Damrauer, J.S., Knijnenburg, T.A., Robertson, A.G., Yau, C., Zhou, W., Berger, A.C., Huang, K.-L., Newberg, J.Y., et al. (2020). Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. Cancer Cell *37*, 639–654.e6.
3. Koga, Y., Song, H., Chalmers, Z.R., Newberg, J., Kim, E., Carrot-Zhang, J., Piou, D., Polak, P., Abdulkadir, S.A., Ziv, E., et al. (2020). Genomic Profiling of Prostate Cancers from Men with African and European Ancestry. Clin. Cancer Res. *26*, 4651–4660.
4. Sinha, S., Mitchell, K.A., Zingone, A., Bowman, E., Sinha, N., Schäffer, A.A., Lee, J.S., Ruppin, E., and Ryan, B.M. (2020). Higher prevalence of homologous recombination deficiency in tumors from African Americans versus European Americans. Nat. Can. (Ott.) *1*, 112–121.
5. Huo, D., Hu, H., Rhie, S.K., Gamazon, E.R., Cherniack, A.D., Liu, J., Yoshimatsu, T.F., Pitt, J.J., Hoadley, K.A., Troester, M., et al. (2017). Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. JAMA Oncol. *3*, 1654–1662.
6. Jiagge, E., Jibril, A.S., Chitale, D., Bensenhaver, J.M., Awuah, B., Hoenerhoff, M., Adjei, E., Bekele, M., Abebe, E., Nathanson, S.D., et al. (2016). Comparative Analysis of Breast Cancer Phenotypes in African American, White American, and West Versus East African patients: Correlation Between African Ancestry and Triple-Negative Breast Cancer. Ann. Surg Oncol. *23*, 3843–3849.
7. Malone, E.R., Oliva, M., Sabatini, P.J.B., Stockley, T.L., and Siu, L.L. (2020). Molecular profiling for precision cancer therapies. Genome Med. *12*, 8.
8. Brown, N.A., and Elenitoba-Johnson, K.S.J. (2020). Enabling Precision Oncology Through Precision Diagnostics. Annu. Rev. Pathol. *15*, 97–121.
9. Mosele, F., Remon, J., Mateo, J., Westphalen, C.B., Barlesi, F., Lolkema, M.P., Normanno, N., Scarpa, A., Robson, M., Meric-Bernstam, F., et al. (2020). Recommendations for the use of next-generation sequencing (NGS) for patients with metastatic cancers: a report from the ESMO Precision Medicine Working Group. Ann. Oncol. *31*, 1491–1505.
10. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
11. Myer, P.A., Lee, J.K., Madison, R.W., Pradhan, K., Newberg, J.Y., Isasi, C.R., Klempner, S.J., Frampton, G.M., Ross, J.S., Venstrom, J.M., et al. (2022). The Genomics of Colorectal Cancer in Populations with African and European Ancestry. Cancer Discov. *12*, 1282–1293.
12. Arora, K., Tran, T.N., Kemel, Y., Mehine, M., Liu, Y.L., Nandakumar, S., Smith, S.A., Brannon, A.R., Ostrovnaya, I., Stopsack, K.H., et al. (2022). Genetic Ancestry Correlates with Somatic Differences in a Real-World Clinical Cancer Sequencing Cohort (Cancer Discov). OF1–OF14.
13. Siegel, R.L., Miller, K.D., Fuchs, H.E., and Jemal, A. (2022). Cancer statistics, 2022. CA A Cancer J. Clin. *72*, 7–33.
14. Giaquinto, A.N., Miller, K.D., Tossas, K.Y., Winn, R.A., Jemal, A., and Siegel, R.L. (2022). Cancer statistics for African American/Black People 2022. CA A Cancer J. Clin. *72*, 202–229.
15. Zakharia, F., Basu, A., Absher, D., Assimes, T.L., Go, A.S., Hlatky, M.A., Iribarren, C., Knowles, J.W., Li, J., Narasimhan, B., et al. (2009). Characterizing the admixed African ancestry of African Americans. Genome Biol. *10*, R141.
16. Cote, M.L., Haddad, R., Edwards, D.J., Atikukke, G., Gadgeel, S., Soubani, A.O., Lonardo, F., Bepler, G., Schwartz, A.G., and Ethier, S.P. (2011). Frequency and type of epidermal growth factor receptor mutations in African Americans with non-small cell lung cancer. J. Thorac. Oncol. *6*, 627–630.
17. Reinersman, J.M., Johnson, M.L., Riely, G.J., Chitale, D.A., Nicastri, A.D., Soff, G.A., Schwartz, A.G., Sima, C.S., Ayalew, G., Lau, C., et al. (2011). Frequency of EGFR and KRAS mutations in lung adenocarcinomas in African Americans. J. Thorac. Oncol. *6*, 28–31.
18. Campbell, J.D., Lathan, C., Sholl, L., Ducar, M., Vega, M., Sunkavalli, A., Lin, L., Hanna, M., Schubert, L., Thorner, A., et al. (2017). Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations. JAMA Oncol. *3*, 801–809.
19. Nassar, A.H., Adib, E., and Kwiatkowski, D.J. (2021). Distribution of KRAS G12C Somatic Mutations across Race, Sex, and Cancer Type. N. Engl. J. Med. *384*, 185–187.
20. Adib, E., Nassar, A.H., Abou Alaiwi, S., Groha, S., Akl, E.W., Sholl, L.M., Michael, K.S., Awad, M.M., Jänne, P.A., Gusev, A., and Kwiatkowski, D.J. (2022). Variation in targetable genomic alterations in non-small cell lung cancer by genetic ancestry, sex, smoking history, and histology. Genome Med. *14*, 39.
21. Krueger, G.G., Papp, K.A., Stough, D.B., Loven, K.H., Gulliver, W.P., and Ellis, C.N.; Alefacept Clinical Study Group (2002). A randomized, double-blind, placebo-controlled phase III study evaluating efficacy and tolerability of 2 courses of alefacept in patients with chronic plaque psoriasis. J. Am. Acad. Dermatol. *47*, 821–833.
22. Swisher, E.M., Lin, K.K., Oza, A.M., Scott, C.L., Giordano, H., Sun, J., Konecny, G.E., Coleman, R.L., Tinker, A.V., O'Malley, D.M., et al. (2017). Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma (ARIEL2 Part 1): an international, multicentre, open-label, phase 2 trial. Lancet Oncol. *18*, 75–87.

23. Nassar, A.H., Adib, E., Abou Alaiwi, S., El Zarif, T., Groha, S., Akl, E.W., Nuzzo, P.V., Mouhieddine, T.H., Perea-Chamblee, T., Taraszka, K., et al. (2022). Ancestry-driven recalibration of tumor mutational burden and disparate clinical outcomes in response to immune checkpoint inhibitors. Cancer Cell *40*, 1161–1172.e5.

24. Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precis. Oncol. *2017*, 1–16.

25. Mahal, B.A., Alshalalfa, M., Kensler, K.H., Chowdhury-Paulino, I., Kantoff, P., Mucci, L.A., Schaeffer, E.M., Spratt, D., Yamoah, K., Nguyen, P.L., and Rebbeck, T.R. (2020). Racial Differences in Genomic Profiling of Prostate Cancer. N. Engl. J. Med. *383*, 1083–1085.

26. Kamran, S.C., Xie, J., Cheung, A.T.M., Mavura, M.Y., Song, H., Palapattu, E.L., Madej, J., Gusev, A., Van Allen, E.M., and Huang, F.W. (2021). Tumor Mutations Across Racial Groups in a Real-World Data Registry. JCO Precis. Oncol. *5*, 1654–1658.

27. Goel, N., Kim, D.Y., Guo, J.A., Zhao, D., Mahal, B.A., and Alshalalfa, M. (2022). Racial Differences in Genomic Profiles of Breast Cancer. JAMA Netw. Open *5*, e220573.

28. Haiman, C.A., Stram, D.O., Wilkens, L.R., Pike, M.C., Kolonel, L.N., Henderson, B.E., and Le Marchand, L. (2006). Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer. N. Engl. J. Med. *354*, 333–342.

29. Aldrich, M.C., Mercaldo, S.F., Sandler, K.L., Blot, W.J., Grogan, E.L., and Blume, J.D. (2019). Evaluation of USPSTF Lung Cancer Screening Guidelines Among African American Adult Smokers. JAMA Oncol. *5*, 1318–1324.

30. Stram, D.O., Park, S.L., Haiman, C.A., Murphy, S.E., Patel, Y., Hecht, S.S., and Le Marchand, L. (2019). Racial/Ethnic Differences in Lung Cancer Incidence in the Multiethnic Cohort Study: An Update. J. Natl. Cancer Inst. *111*, 811–819.

31. Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. *23*, 703–713.

32. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

33. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. *47*, D941–D947.

34. Alexander, D.H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinf. *12*, 246.

35. Li, H. (2010). Aligning New-Sequencing Reads by BWA (Broad Institute).

36. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

37. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

38. Belsare, S., Levy-Sakin, M., Mostovoy, Y., Durinck, S., Chaudhuri, S., Xiao, M., Peterson, A.S., Kwok, P.-Y., Seshagiri, S., and Wall, J.D. (2019). Evaluating the quality of the 1000 genomes project data. BMC Genom. *20*, 620.

39. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

40. Frampton, G.M., Fichtenholtz, A., Otto, G.A., Wang, K., Downing, S.R., He, J., Schnall-Levin, M., White, J., Sanford, E.M., An, P., et al. (2013). Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat. Biotechnol. *31*, 1023–1031.

41. He, J., Abdel-Wahab, O., Nahas, M.K., Wang, K., Rampal, R.K., Intlekofer, A.M., Patel, J., Krivstov, A., Frampton, G.M., Young, L.E., et al. (2016). Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. Blood *127*, 3004–3014.

42. Milbury, C.A., Creeden, J., Yip, W.-K., Smith, D.L., Pattani, V., Maxwell, K., Sawchyn, B., Gjoerup, O., Meng, W., Skoletsky, J., et al. (2022). Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. PLoS One *17*, e0264138.

43. Sun, J.X., He, Y., Sanford, E., Montesion, M., Frampton, G.M., Vignot, S., Soria, J.-C., Ross, J.S., Miller, V.A., Stephens, P.J., et al. (2018). A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. PLoS Comput. Biol. *14*, e1005965.

44. Chalmers, Z.R., Connelly, C.F., Fabrizio, D., Gay, L., Ali, S.M., Ennis, R., Schrock, A., Campbell, B., Shlien, A., Chmielecki, J., et al. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med. *9*, 34.

45. Trabucco, S.E., Gowen, K., Maund, S.L., Sanford, E., Fabrizio, D.A., Hall, M.J., Yakirevich, E., Gregg, J.P., Stephens, P.J., Frampton, G.M., et al. (2019). A Novel Next-Generation Sequencing Approach to Detecting Microsatellite Instability and Pan-Tumor Characterization of 1000 Microsatellite Instability-High Cases in 67,000 Patient Samples. J. Mol. Diagn. *21*, 1053–1066.

46. Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. *23*, 703–713.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
1000 Genomes project data	The International Genome Sample Resource	RRID: SCR_006828
MSK-IMPACT	Zehir et al.[31]	https://cbioportal.mskcc.org/
ExAC	Lek et al.[32]	https://gnomad.broadinstitute.org/
COSMIC	Tate et al.[33]	https://cancer.sanger.ac.uk
Zip code-based income inference	IRS Revenue Procedure 2018-57 Missouri Census Data Center	https://www.ncsha.org/resource/rev-proc-18-57/ https://mcdc.missouri.edu/
Software and algorithms		
ADMIXTURE	Alexander et al.[34]	https://dalexander.github.io/admixture/
BWA	Li and Durbin[35]	http://bio-bwa.sourceforge.net/
Picard	Broad Institute	http://broadinstitute.github.io/picard/
SAMtools	Li et al.[36]	http://www.htslib.org/
GATK	McKenna et al.[37]	https://gatk.broadinstitute.org/
OncoKB	Chakravarty et al.[24]	https://www.oncokb.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jian Carrot-Zhang (carrotj@mskcc.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Consented data that can be released are included in the article and its supplementary files. Tumor mutations and associated clinical data for 50,000 patients in this study are available through AACR Project GENIE (v10.1 public cohort). Patients did not consent for the release of underlying raw genomic sequence data. Academic researchers can gain access to the data by contacting the corresponding authors. For the Foundation Medicine, Inc. data, you and your institution will be required to execute a data transfer agreement. For further questions please reach out to the compliance department at Foundation Medicine, Inc., Cambridge, MA (compliance@foundationmedicine.com).

This paper does not report original code. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Approval for this study, including a waiver of informed consent and Health Insurance Portability and Accountability Act waiver of authorization, was obtained from the Western Institutional Review Board (Protocol #20152817). The Institutional Review Board granted a waiver of informed consent under 45 CFR § 46.116 based on review and determination that this research meets the following requirements: (i) the research involves no more than minimal risk to the subjects; (ii) the research could not practicably be carried out without the requested waiver; (iii) the waiver will not adversely affect the rights and welfare of the subjects. The validation cohort was approved by the Memorial Sloan Kettering Cancer Center Institutional review board and all patients provided written informed consent for tumor sequencing and review of patient medical records for detailed demographic, pathologic, and treatment information (NCT01775072). All samples were obtained during routine clinical care.

METHOD DETAILS

Ancestry assignment

For the discovery cohort, genetic ancestry was inferred on over 333,000 patient tumor samples using over >40,000 germline single nucleotide polymorphisms (SNPs) included in the targeted gene panel sequencing.[2] The ancestry inference method was described in

previous studies.[2,3,11] We trained an ancestry caller, using at least 39,540 SNPs cataloged by both the 1000 Genomes Project and by our sequencing assays.[2,38] These SNPs were projected using principal component analysis, and the top *N* resulting features were used to train a random forest classifier (*N*=10). Individuals were first classified into one of five inferred population ancestry groups: AFR, EUR, AMR, SAS, or EAS. Subclassification of patients of AFR ancestry was performed to infer one of the following three subgroups: WAFR, EAFR, and AAFR, and samples that were not classified as one of the three sub AFR groups were removed from further analysis.

Tenfold cross validation was performed on the 1000 Genomes samples to assess classifier performance. In total, there are 2,504 samples in the 1000 Genomes dataset, comprising 661 AFR, 347 AMR, 504 EAS, 503 EUR, and 489 SAS samples. Each fold consists of approximately 250 randomly sampled (without replacement) individuals. For each fold tested, a classifier was trained on the other nine folds, and then applied to the testing fold. This produced calls for each sample in the validation set, which were then compared with the ground truth population labels. Due to the slight imbalance in population representation, precision and recall were determined for each population and then a single class-weighted precision and class-weighted recall metric was calculated from the mean of these:

TP_p = Call matches ground truth for a population, p
FP_p = Number of calls erroneously attributed to a population, p
FN_p = Number of calls erroneously not attributed to a population, p

$$Precision_p = TP_p / (TP_p + FP_p)$$

$$Recall_p = TP_p / (TP_p + FN_p)$$

$$Precision = (Precision_{AFR} + Precision_{AMR} + Precision_{EAS} + Precision_{EUR} + Precision_{SAS}) / N_p$$

$$Recall = (Recall_{AFR} + Recall_{AMR} + Recall_{EAS} + Recall_{EUR} + Recall_{SAS}) / N_p$$

where N_p is 5, for the five population groups.

For the validation cohort, the same ancestry inference approach was used as previously described.[12] Ancestry calls were made using >3,000 SNPs in the MSK-IMPACT matched normal sequencing data.[12] The ADMIXTURE tool[39] was used to estimate the proportion of each ancestry for both the discovery and the validation cohort, using targeted SNPs in each cohort. The discovery cohort and the validation cohort used the 1000 Genomes samples as reference and used k=5 to estimate the relative AFR, AMR, EAS, SAS and EUR ancestry admixture. Then, YRI, GWD, MSL, LWK, ESN, CEU, TSI, GBR, and IBS samples in the 1000 Genomes were also used to infer the relative WAFR, EAFR, and EUR ancestry admixture in the discovery cohort with k=3.

Genomic profiling and alteration detection

Targeted gene panel sequencing was performed in a Clinical Laboratory Improvement Amendments-certified, College of American Pathologists-accredited laboratory on de-identified, research-consented, formalin-fixed paraffin-embedded tumor specimens obtained over the course of routine cancer care in the USA, which contained a mixture of tumor ($\geq$ 20%) and normal tissue.[40–42] At least 50 ng of DNA per specimen was isolated and sequenced to high, uniform coverage (mean >500×), as previously described.[40–42] For samples run with the FoundationOne® Heme assay, at least 300 ng of RNA was also isolated and converted to complementary DNA, and sequenced to >3 million on-target pairs to aid in rearrangement detection.[40–42]

Sequence analysis methods and validation used in this study have been described previously.[40–42] Reads were mapped to the hg19 reference genome using BWA aligner.[35] Sequencing metrics and duplicate reads were removed using Picard (https://github.com/broadinstitute/picard) and SAMtools.[36] Local alignment optimization was performed using GATK 1.0.4705.[37] Variant calls were limited to targeted genomic regions.[40–42] Base substitution detection was performed using the Bayesian methodology, which enables the detection of novel mutations at low mutant allele frequency and increased sensitivity for mutations at hot-spot sites through the incorporation of tissue-specific prior expectations.[40–42] To detect short insertions or deletions (indels), *de novo* local assembly in each targeted exon was performed using the De Bruijn approach.[40–42] After read pairs were collected and decomposed, the statistical support for competing haplotypes was evaluated and candidate indels were aligned against the reference genome. Filtering of indel candidates was carried out as described for base substitutions.[40–42]

Alterations were classified as known or likely pathogenic based on COSMIC[33] association, characterization in the scientific literature, and knowledge about the gene affected (i.e., truncations in known tumor suppressor genes). Copy number amplifications in oncogenes and homozygous copy number deletions of tumor suppressors were also classified as known or likely pathogenic. Additionally, rearrangements that predicted to activate oncogenes or inactivate tumor suppressors and recurrent fusions were annotated as known or likely pathogenic. Remaining alterations were classified as unknown and excluded from gene-specific analyses.

For TMB and mutational signature analysis, known or likely pathogenic alterations were not counted toward the total. Additionally, alterations predicted to be germline by the somatic-germline-zygosity algorithm[43] were not included. Alterations that were recurrently predicted to be germline in any ancestry were also not counted. Known germline alterations in dbSNP were not counted.

Germline alterations occurring more than twice in the ExAC database[32] were not included. TMB was defined as the number of somatic targeted coding region short variants (base substitution or indel) per megabase of genome examined. Patients were stratified into TMB bins of either TMB-high (TMB $\geq$ 10 mutations [mut]/megabase [Mb]) or TMB-low (TMB<10 mut/Mb). Microsatellite instability calling was determined on up to 114 intronic homopolymer repeat loci with a minimum of 250x median depth. Means and variances of repeat lengths across reads were used as input into principal component analysis. The first principal component was used as an MSI-score to determine MSI-high status, MSI-ambiguous status, or microsatellite stable (MSS) status.[44,45] Trinucleotide mutational signatures were generated as described previously,[46] where the 96 single-base substitution COSMIC reference (version 2) was used to perform the decomposition. The smoking signature was defined as the sums of COSMIC signature 4 and 29 coefficient weights. Signatures were called only for samples with at least 10 assessable mutations and if the score was $\geq$0.4.

Gene amplifications and homozygous deletions were detected by comparing complete chromosomal copy number maps to reference process-matched, normal control samples.[40–42] Log-ratio profiles were generated by normalizing the sequence coverage obtained at all exons and genome-wide SNPs to the process-matched normal control and then GC bias corrected. Profiles were then segmented and interpreted by using allele frequencies of sequenced SNPs to estimate tumor purity and copy number at each segment. Gene fusions and rearrangements were detected by analysis of chimeric read pairs from DNA (pairs mapping to different chromosomes or > 10 kilobase pairs apart) and RNA (pairs mapping to different genes or genomic loci > 10 kilobase pairs apart), when available.[40–42] Percent genome-wide loss of heterozygosity (gLOH) was calculated from 22 autosomal chromosomes using the genome-wide copy number profiles, as well as minor allele fractions of over 3,500 sequenced SNPs. Loss of heterozygosity was called if the estimated copy number was not 0 but the minor allele count was 0 at a given segment. Segments with at least 90% loss of heterozygosity across a chromosome or an arm of a chromosome and segments for which loss of heterozygosity inference was ambiguous were not included in the percent genome-wide loss of heterozygosity calculations.

QUANTIFICATION AND STATISTICAL ANALYSIS

Identification of ancestry association

Multivariate logistic or linear regression was used to test the association of genomic alterations with AFR ancestry percentage in each individual for each disease - controlling for clinical and technical covariates when applicable (age, sex, TMB, panel version, subtype, disease stage). Only known, likely functional genomic alterations or OncoKB annotated mutations (sop.oncokb.org) in genes of interest were included in the analyses. False discovery rates (FDRs) were calculated using the Benjamini–Hochberg approach. The effect of smoking exposure on driver alteration modified by AFR ancestry was analyzed using:

$$\text{Logit (S)} = \beta 0 + \beta 1\text{A} + \beta 2\text{E} + \beta 3\text{C} + \beta 4(\text{A x E})$$

where S is the LUAD driver gene (*TP53*, *KRAS*, *EGFR*, *BRAF*, *ALK* fusion and *ROS1* fusion), A is AFR ancestry, E is smoking-related environmental exposure (smoking mutational signature or self-reported smoking status), C is a set of covariates when applicable (age, sex, stage), and A x E is the ancestry-environment interaction term. FDR correction was applied for the P values of the interaction term. Statistical tests were performed using R version 3.4 or Python version 2.7.

Clinical data analysis

Clinical variables in the MSK-IMPACT cohort were extracted from linked electronic health records. Overall survival (OS) was measured from the time of sequencing to death and was censored at the last time the patient was known to be alive. Inferred income was assigned based on the 2019 marginal tax rate (https://www.ncsha.org/resource/rev-proc-18-57/) corresponding to the patient zip code (https://mcdc.missouri.edu/). Patients living in a zip code with a median income equal to or exceeding $84,200 (tax rate for heads of household $\geq$24%) were assigned to the 'Tax Tier: High-Income' category while patients in a zip code below that threshold were assigned to the 'Tax Tier: Low-Income' category. Multivariable Cox proportional hazards models were used to associate the OS with ancestry as a continuous variable, and ancestry-associated alterations, accounting for covariates including age, sex, histologic subtype, disease stage, fraction of genome altered, average BMI, inferred income, and the most recent insurance type (medicaid or non-medicaid). The interaction of AFR ancestry and biomarker on survival was analyzed using:

$$\text{Cox (O)} = \beta_0 + \beta_1\text{S} + \beta_2\text{A} + \beta_3\text{C} + \beta_4(\text{A x S})$$

where O is OS, A is ancestry, S is the alteration, C is a set of covariates, and A x S is the ancestry-somatic interaction term, which is our figure of interest. This interaction term (β4) indicates the difference in the effect of S on OS between ancestries within mutant vs. wildtype samples. FDR correction was applied for the P values of the interaction term.