Rhead *et al. Genome Medicine* (2024) 16:99
https://doi.org/10.1186/s13073-024-01373-w

Genome Medicine

RESEARCH

Open Access

Association of genetic ancestry with molecular tumor profiles in colorectal cancer

Brooke Rhead[1†], David M. Hein[2†], Yannick Pouliot[1], Justin Guinney[1], Francisco M. De La Vega[1,3*] and Nina N. Sanford[2*]

Abstract

Background There are known disparities in incidence and outcomes of colorectal cancer (CRC) by race and ethnicity. Some of these disparities may be mediated by molecular changes in tumors that occur at different rates across populations. Genetic ancestry is a measure complementary to race and ethnicity that can overcome missing data issues and better capture genetic similarity in admixed populations. We aimed to identify somatic mutations and tumor gene expression differences associated with both genetic ancestry and imputed race and ethnicity.

Methods Sequencing was performed with the Tempus xT NGS 648-gene panel and whole exome capture RNA-Seq for 8454 primarily late-stage CRC patients. Genetic ancestry proportions for five continental groups—Africa (AFR), American indigenous (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS)—were estimated using ancestry informative markers. To address data gaps, race and ethnicity categories were imputed, resulting in assignments for 952 Hispanic/Latino, 420 non-Hispanic (NH) Asian, 1061 NH Black, and 5763 NH White individuals. We assessed association of genetic ancestry proportions and imputed race and ethnicity categories with somatic mutations in relevant CRC genes and in 2608 expression profiles, as well as 1957 consensus molecular subtypes (CMS).

Results Increased AFR ancestry was associated with higher odds of somatic mutations in *APC*, *KRAS*, and *PIK3CA* and lower odds of *BRAF* mutations. Additionally, increased EAS ancestry was associated with lower odds of mutations in *KRAS*, EUR with higher odds in *BRAF*, and the Hispanic/Latino category with lower odds in *BRAF*. Greater AFR ancestry and the NH Black category were associated with higher rates of CMS3, while a higher proportion of Hispanic/Latino patients exhibited indeterminate CMS classifications.

Conclusions Molecular differences in CRC tumor mutation frequencies and gene expression that may underlie observed differences by race and ethnicity were identified. The association of AFR ancestry with increased *KRAS* mutations aligns with higher CMS3 subtype rates in NH Black patients. The increase of indeterminate CMS in Hispanic/Latino patients suggests that subtype classification methods could benefit from enhanced patient diversity.

[†]Brooke Rhead and David M. Hein contributed equally to this work and share first authorship.

*Correspondence:
Francisco M. De La Vega
francisco.delavega@stanford.edu
Nina N. Sanford
nina.sanford@utsouthwestern.edu
Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

Overall incidence and mortality of colorectal cancer (CRC) has declined over the last several decades due to a combination of risk reduction, early detection, and advancements in therapy [1]. However, there has been a growing burden of CRC among young adults and persistent disparities in outcomes by race and ethnicity across all ages [2]. As such, improved CRC outcomes are not equally realized across demographics in the United States.

The rising incidence of CRC among adults aged < 50 years, termed early onset CRC (EOCRC), has garnered significant attention by patients, media, and clinicians. Patients with EOCRC typically have delayed presentation, leading to more advanced disease at time of diagnosis [3]. To date, studies have not demonstrated consistent, clinically relevant molecular differences in early versus average onset CRC (AOCRC) [4–6]. As such, the cause for increasing incidence of EOCRC is largely attributed to potential environmental and behavioral components, with specific factors yet to be elucidated [3].

Racial and ethnic differences in CRC outcomes are also multifactorial in etiology. Longstanding disparities in access to care have disproportionately affected Black populations who have the worst CRC outcomes, regardless of clinical factors such as age or stage at diagnosis [7–9]. Prior studies have also demonstrated molecular differences in CRC by race and ethnicity with predictive and prognostic implications, including increased prevalence of *KRAS* mutations among Black patients [8, 10–12]. Most of these studies use self-reported or observed race and ethnicity categories. In healthcare and clinico-genomic databases, a high proportion of this information (30–70%) is often missing, and when it is present, it may be based on clinician observations rather than self-reported by patients [13–15]. Furthermore, race and ethnicity categories may not capture shared ancestry well in highly admixed groups such as Black and Hispanic/Latino patients [16]. In contrast, genetic ancestry, assessed via a patient's sequencing or genotyping data, can potentially better capture genotypic profiles associated with risk, though it is important to understand that genetic ancestry is also associated with environmental risks [17].

Given disparities in incidence and outcome of CRC by race, ethnicity, and age, along with the limitations of traditionally used race and ethnicity categories based on the US government's Office of Management and Budget standard [18], we examined whether genetic ancestry proportions were associated with patterns of molecular alterations in CRC using a large, cohort from the Tempus clinico-genomic database. This database compiles multimodal genomic and clinical data from cancer patient care and can facilitate molecular pathological epidemiology studies aimed at exploring the interplay between individual factors such as clinical measurements, genetic ancestry, or race/ethnicity, and molecular tumor traits, environmental influences, and clinical outcomes [19]. This is a convenience sample, which, despite its scale and diversity surpassing that of research and clinical trial studies, may still harbor unknown ascertainment biases [20]. Such factors could potentially affect the generalizability of our findings. To address the missingness of race and ethnicity data common in this dataset, we imputed these categories from genetic ancestry [21]. We then evaluated associations with the imputed categories, both to compare to our genetic ancestry proportion findings and to prior research using self-reported categories. Furthermore, we assessed whether race and ethnicity associations were different in AOCRC versus EOCRC, or by primary tumor site.

Methods

Patient cohort

Genomic and clinical data of 8454 patients diagnosed with CRC were obtained from the Tempus database, which includes de-identified genomic and clinical data from cancer patients that underwent tumor profiling as part of their healthcare. Selection criteria included tumor profiling with the Tempus xT assay (v2–v4) from 2018 to 2022. Briefly, the assay is a targeted panel that detects single nucleotide variants, insertions and/or deletions, and copy number variants in 598–648 genes, as well as chromosomal rearrangements in 22 genes with high sensitivity and specificity. A subset of those patients with sufficient tumor sample material had additional whole exome RNA sequencing [22]. Available demographic information included patient age at date of specimen collection, age at diagnosis, gender, stated (i.e., either self-reported or observed) race and ethnicity, and smoking status. Primary tumor site, clinical details such as tumor grade, microsatellite instability (MSI) status, tumor mutational burden count (TMB, number of mutations/megabase), and sequenced tissue site were included. All analyses were performed using de-identified data.

Patient characteristics are summarized in Table 1. Among the cohort of 8454 CRC patients, 5169 (61%) had a matched normal tissue sample and 2745 (32%) had RNA sequencing performed on the tumor sample. The median age was 60.7 years (IQR 51.1–69.6) with 1792 (25.6% of patients with available diagnosis age) diagnosed under the age of 50 (i.e., with EOCRC) and 7997 (94.6%) with microsatellite stable (MSS) disease. Most patients had advanced disease, with 4254 (81% of those with known stage) diagnosed with stage IV disease. Onset age group differed by imputed race and ethnicity category

Table 1 Patient characteristics by imputed race and ethnicity category. Columns contain *n* (%) for categorical variables or median (IQR) for continuous variables. *p* values for categorical variables with any expected cell count < 5 are from a Fisher's exact test with a simulated *p* value based on 2000 replicates; *p* values for categorical variables with all expected cell counts ≥ 5 are from a Pearson's chi-squared test; and *p* values for continuous variables are from a Kruskal-Wallis rank sum test

Characteristic	Complex, *N* = 2581	Hispanic/Latino, *N* = 9521	NH Asian, *N* = 4201	NH Black, *N* = 10,611	NH White, *N* = 57,631	*p* value
Stated race						< 0.001
White	95 (78%)	223 (57%)	9 (4.1%)	23 (3.4%)	3268 (97%)	
American Indian or Alaska Native	3 (2.5%)	29 (7.4%)	3 (1.4%)	0 (0%)	2 (< 0.1%)	
Asian	10 (8.2%)	1 (0.3%)	181 (83%)	0 (0%)	2 (< 0.1%)	
Black or African American	0 (0%)	6 (1.5%)	0 (0%)	634 (94%)	9 (0.3%)	
Native Hawaiian or Other Pacific Islander	2 (1.6%)	1 (0.3%)	4 (1.8%)	0 (0%)	1 (< 0.1%)	
Other race	12 (9.8%)	129 (33%)	19 (8.8%)	20 (3.0%)	75 (2.2%)	
Race not stated	0 (0%)	1 (0.3%)	1 (0.5%)	0 (0%)	6 (0.2%)	
Unknown	136	562	203	384	2400	
Stated ethnicity						< 0.001
Not Hispanic or Latino	64 (75%)	67 (14%)	120 (99%)	277 (95%)	1717 (98%)	
Hispanic or Latino	21 (25%)	414 (86%)	1 (0.8%)	14 (4.8%)	37 (2.1%)	
Unknown	173	471	299	770	4009	
Age at specimen collection	59 (50, 69)	56 (47, 66)	60 (50, 68)	60 (50, 68)	62 (52, 70)	< 0.001
Unknown	0	3	0	1	10	
Age at onset	57 (48, 67)	55 (45, 64)	58 (49, 66)	58 (49, 67)	60 (51, 69)	< 0.001
Unknown	47	142	64	185	1003	
Onset age group						< 0.001
AOCRC	153 (73%)	514 (63%)	255 (72%)	635 (72%)	3664 (77%)	
EOCRC	58 (27%)	296 (37%)	101 (28%)	241 (28%)	1096 (23%)	
Unknown	47	142	64	185	1003	
Gender						0.4
Female	115 (45%)	420 (44%)	179 (43%)	487 (46%)	2455 (43%)	
Male	143 (55%)	528 (56%)	239 (57%)	572 (54%)	3287 (57%)	
Unknown	0	4	2	2	21	
xT assay version						0.3
xT.v2	36 (14%)	120 (13%)	68 (16%)	149 (14%)	857 (15%)	
xT.v3	44 (17%)	194 (20%)	64 (15%)	204 (19%)	1076 (19%)	
xT.v4	178 (69%)	638 (67%)	288 (69%)	708 (67%)	3830 (66%)	
Smoking status						< 0.001
Never smoker	86 (55%)	375 (60%)	184 (67%)	392 (55%)	2007 (51%)	
Ever smoker	70 (45%)	255 (40%)	89 (33%)	323 (45%)	1893 (49%)	
Unknown	102	322	147	346	1863	
Cancer stage						0.019
Stage 1	2 (1.3%)	5 (0.9%)	1 (0.4%)	7 (1.0%)	23 (0.6%)	
Stage 2	11 (7.1%)	24 (4.2%)	7 (2.8%)	32 (4.6%)	163 (4.6%)	
Stage 3	21 (14%)	107 (19%)	40 (16%)	89 (13%)	444 (12%)	
Stage 4	120 (78%)	439 (76%)	202 (81%)	568 (82%)	2925 (82%)	
Unknown	104	377	170	365	2208	
Tumor grade						0.002
Low	7 (4.5%)	78 (12%)	30 (10%)	56 (7.9%)	361 (9.5%)	
Medium	110 (70%)	474 (71%)	199 (69%)	511 (72%)	2549 (67%)	
High	40 (25%)	117 (17%)	61 (21%)	144 (20%)	899 (24%)	
Unknown	101	283	130	350	1954	

Table 1 (continued)

Characteristic	Complex, $N=2581$	Hispanic/Latino, $N=9521$	NH Asian, $N=4201$	NH Black, $N=10,611$	NH White, $N=57,631$	*p* value
MSI status						0.062
Low/stable	247 (96%)	896 (94%)	405 (96%)	1017 (96%)	5432 (94%)	
High	10 (3.9%)	56 (5.9%)	15 (3.6%)	42 (4.0%)	322 (5.6%)	
Unknown	1	0	0	2	9	
TMB count (mutations/Mb)	3 (2, 5)	3 (2, 5)	3 (2, 5)	4 (2, 6)	3 (2, 5)	< 0.001
Unknown	0	0	1	2	4	
Tumor/normal tissue status						0.022
Tumor and normal	152 (59%)	626 (66%)	246 (59%)	659 (62%)	3486 (61%)	
Tumor only	105 (41%)	326 (34%)	174 (41%)	401 (38%)	2270 (39%)	
Unknown	1	0	0	1	7	
Cancer primary site						< 0.001
Left colon	34 (13%)	98 (10%)	52 (12%)	88 (8.3%)	524 (9.1%)	
Not specified	153 (59%)	609 (64%)	253 (60%)	712 (67%)	3823 (66%)	
Rectum	46 (18%)	161 (17%)	72 (17%)	128 (12%)	850 (15%)	
Right colon	25 (9.7%)	84 (8.8%)	43 (10%)	133 (13%)	566 (9.8%)	
Sequenced tissue site						< 0.001
CRC (primary)	137 (53%)	554 (58%)	211 (50%)	556 (53%)	2841 (49%)	
Liver	47 (18%)	185 (19%)	75 (18%)	242 (23%)	1270 (22%)	
Lung	11 (4.3%)	43 (4.5%)	39 (9.3%)	41 (3.9%)	347 (6.0%)	
Other	62 (24%)	168 (18%)	95 (23%)	220 (21%)	1285 (22%)	
Unknown	1	2	0	2	20	

($p < 0.001$, Table 1). The Hispanic/Latino category had the highest proportion of EOCRC (37%), while NH White had the lowest proportion (23%). See Additional file 1: Tables S2 and S3 for patient characteristics stratified by MSI status, and Additional file 1: Tables S4 and S5 for patient characteristics stratified by onset age group.

CRC-relevant genes and mutation types

Genes relevant to CRC were identified from the following sources: 187 genes belonging to 10 oncogenic signaling pathways reported by Sanchez-Vega et al. [23], 72 genes predicted by the Integrative OncoGenomics pipeline to be CRC drivers (IntOGen, release date 2020.02.01) [24], 15 genes associated with hereditary colorectal cancer syndromes (Lynch, Li Fraumeni, and polyposis syndromes) for which germline variants are reportable in the Tempus xT assay, and 22 genes that were investigated in a previous CRC study that utilized Tempus data [10, 22]. Of these genes, 137 are included in the Tempus xT assay gene panels (v2–v4).

Different mutation types were evaluated: (1) protein-altering somatic mutations, defined as single or multiple nucleotide mutations, short insertions or deletions (≤ 50 bp), and other changes that impact protein structure or splice sites (Sequence Ontology, SO:0001818), (2) somatic copy number alterations (SCNAs), defined as structural insertions or deletions greater than 500 bp in size, and (3) actionable mutations, defined in our study as protein-altering mutations with an OncoKB Therapeutic Level of Evidence V2 designation of therapeutic level 1 or 2, or resistance level R1, irrespective of the type of solid cancer [25]. For protein-altering mutations, only patients with matched normal tissues were included in analyses due to the potential for germline variants to be misclassified as somatic when normal tissue is unavailable. Patients without matched normal tissues were included in analyses of SCNAs and actionable mutations. We required a prevalence of at least 1% (and minimum 10 patients) for a specific mutation type to include a gene for evaluation.

Determination of genetic ancestry

Genetic ancestry proportions were estimated using a supervised global genetic ancestry estimation algorithm [26]. An R script implementation is available at DOI 10.7303/syn4877977. Proportions for five continental ancestry groups—Africa (AFR), American indigenous (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS)—were calculated using 654 ancestry informative markers (AIMs) that overlap targeted regions of

the Tempus xT NGS assay [22, 27]. Reference allele frequency data for the AIMs was obtained from the 1000 Genomes Project [28], The Human Genome Diversity Project [29], and the Simons Genome Diversity Project databases [30]. AMR allele frequencies were derived from 22 SDP and 49 HGDP samples encompassing indigenous populations from Argentina, Brazil, Mexico, Colombia, and Peru. The accuracy of our methods was evaluated using published ancestry proportions determined using the gold standard method, RFMix [31], on whole-genome sequencing data from the Pan-Cancer Analysis of Whole Genomes Project (PCAWG) [32], available at DOI 10.7303/syn4877977, and on admixed population samples from the 1000 Genomes Project. We calculated the mean squared errors (MSE) for comparisons across each of the continental groups and computed an average MSE of 0.121 and 0.0141 for the PCAWG and 1000 Genomes projects samples, respectively (cf. Additional file 1: Supplementary Methods, Table S1, and Fig. S1). Normal specimens were used to determine genotypes at the AIMs when available; otherwise, tumor data were used.

Imputation of race and ethnicity categories

To overcome missingness of stated race and ethnicity in our clinico-genomic data (cf. Table 1), imputation of mutually exclusive race and ethnicity categories from genetic ancestry proportions were estimated using a set of heuristics derived from admixture proportions reported in the literature for Black and Hispanic/Latino groups in the United States [33], using a method we previously published [21]. Four categories were defined, non-Hispanic (NH) Asian, NH Black, Hispanic/Latino, and NH White, with patients remaining unclassified under our heuristics termed "complex." The published assessment of the sensitivity and specificity of our imputation method demonstrated high accuracy in our data source (correct rate of 96% and weighted error of 0.9%; no-call rate ~ 3%), enabling us to use this data for comparisons across categories with all patients [21].

Association between genetic ancestry or race and ethnicity category and somatic mutations

Tests were stratified by microsatellite instability (MSI) status as determined by the Tempus xT algorithm. Likelihood ratio tests (LRTs) were used to identify genes in which the presence of somatic mutation was associated with genetic ancestry proportions or race and ethnicity imputed categories. For each gene, a multivariable logistic regression model that included somatic mutation (presence/absence) or copy number alteration as the dependent variable and ancestry proportions, assay version, gender, and age at sample collection as independent variables (full model) was compared to a nested model that excluded ancestry proportions. LRT p values were corrected for the number of genes tested within each somatic mutation type using the Benjamini-Hochberg method. For genes with significant LRT p values, specific genetic ancestry proportion associations (AMR, AFR, EAS, EUR, or SAS) were identified in the full model (any uncorrected coefficient $p < 0.05$ considered significant). Because TMB can vary with race and ethnicity category and could potentially explain significant results from the full model, we repeated the tests within each MSI category, including natural log-transformed TMB as a continuous covariate.

In order to include all five genetic ancestry proportions in the same model, so that each ancestry association was adjusted for the remaining four ancestries while also properly accounting for data compositionality, proportions were first transformed into an isometric log ratio (ILR) representation ("pivot coordinates") using the pivotCoord function in the robCompositions R package [34]. Analyses were then repeated using imputed race and ethnicity categories in place of ancestry proportions, with the "complex" category excluded from further analyses, and the NH White group used as the reference category. Odds ratios (ORs) and 95% confidence intervals were estimated from the full models. Complete case analysis was utilized in all regression models.

Sensitivity analyses were conducted to test whether adjusting for additional covariates affected the associations found in the main analyses. The following variables were individually tested in logistic regression models: sequenced tissue site (CRC, liver, lung, or other), cancer primary site (left colon, not specified, rectum, or right colon), cancer stage (1, 2, 3, or 4), tumor grade (low, medium, or high), high tumor mutational burden (defined as ≥ 10 mutations/megabase as per KEYNOTE-158 study [35]), age at onset (in place of age at collection), cancer primary histology (by restricting to patients with adenocarcinoma only), and smoking status (ever smoker or never smoker). Additional sensitivity tests for cancer primary site were conducted by (1) excluding cases with "not specified" cancer primary site and not adjusting for any additional variables, for the sake of comparison to the next three tests; (2) by excluding cases with "not specified" cancer primary site and adjusting for the remaining categories (left colon, rectum, or right colon); (3) by excluding cases with "not specified" cancer primary site and adjusting for site categorized as left colon/rectum or right colon; and (4) by excluding cases with "not specified" cancer primary site and adjusting for site categorized as colon or rectum.

In addition to the main analyses, we looked for imputed race and ethnicity category associations that

differed by age of diagnosis (EOCRC vs. AOCRC), or by cancer primary site sidedness (colon vs. rectum and left colon/rectum vs. right colon) among microsatellite stable (MSS) patients. To identify such associations, we first conducted LRTs with logistic models similar to those in the main analyses, but with added indicator variables for age of diagnosis (or cancer primary site) in both the full and nested models, and an interaction term for age of diagnosis (or cancer primary site) and imputed race and ethnicity category in the full model. LRT p values were corrected for multiple hypotheses using the Benjamini-Hochberg method. For any genes where evidence of interaction was identified by the LRT, interaction terms in the full model with $p < 0.05$ identified specific race and ethnicity categories with interaction effects. The full models were used to estimate ORs and 95% confidence intervals.

Because patients with MSI-H tumors are often candidates for immunotherapy, even with non-metastatic cancer (NCCN Guidelines version 2.20240), we wished to assess variation in TMB among this cohort [36]. Among patients with MSI-high status, the Kruskal-Wallis test was used to assess whether there were differences in TMB by age at diagnosis (AOCRC vs. EOCRC), by imputed race and ethnicity alone, and by imputed race and ethnicity stratified by age at diagnosis.

Differences in cohort characteristics by imputed race and ethnicity category

Differences in cohort characteristics among imputed race and ethnicity categories were assessed using the R package *gtsummary* [37]. Fisher's exact test for count data with simulated p value (based on 2000 replicates) was used for categorical variables with any expected cell count < 5, Pearson's chi-squared test was used for categorical variables with all expected cell counts ≥ 5, and the Kruskal-Wallis rank sum test was used for continuous variables.

Gene expression data exploration and preparation

Tempus xT RNA-Seq raw sequencing data were processed with Kallisto to quantify transcript abundances as previously described [38]. Raw transcript counts were filtered to a minimum of 10 counts in 5% of samples and a variance stabilized transform (VST, DESeq2) was applied [39]. Batch effects due to assay version were assessed with principal component analysis (PCA) and removed with LIMMA via linear modeling (removeBatchEffect) [40]. PCA plots labeled by grade, MSI status, tissue site, and clinical stage were then generated and inspected for the presence of clustering and used to inform subsetting of patients for separate downstream testing. Further variable selection for multivariable analyses was then performed on each subset. First, variables with more than 25% missing data were removed from consideration. Next, within each subset, PCA plots were again generated for remaining variables to assess their relationship with gene expression. Subsequent differential expression (DE) testing and gene set analyses were performed on data subsets individually using only the covariates appropriate for each subset.

PCA plots of RNA counts demonstrated that after batch correction, tissue site was the primary driver of variation (Additional file 1: Fig. S2). Therefore, we restricted our analyses to liver, colon, and rectum samples (with liver assessed separately from colon/rectum) given small numbers in other metastatic sites (Table 1). Clinical stage was missing for 37% of patients with RNA-Seq results from the colon/rectum or liver, thus was not considered further. PCA plots were generated and labeled by MSI-status and tumor grade, and tumor tissue site for the colon/rectum subset. Given the small number of patients with MSI-high tumors (Additional file 1: Table S3) and differences in gene expression by MSI status (Additional file 1: Fig. S3), MSI-high tumors were excluded from this analysis. There was notable clustering of patients with missing tumor grade for the colon/rectum group with 21% of patients missing grade (Additional file 1: Fig. S3). Given the presence of strong clustering by grade, and grade likely missing not at random, a missing indicator approach was used. This method has been shown to produce an almost unbiased result while preserving power lost under complete case analysis [41]. Final variables included for multivariable analysis were tumor grade, gender, early versus average onset, colon vs. rectum tumor tissue site (not present for liver subgroup), and either imputed race and ethnicity categories or pivot coordinates for genetic ancestry proportions.

Gene set analysis workflow 1: GSVA

Because gene set testing approaches test somewhat different hypotheses, we performed gene set analysis in the Hallmark and C2 Biocarta gene sets (342 total) from MSigDB using two distinct workflows [42–44]. GSVA is a method that evaluates the expression of genes within a gene set relative to those outside of the set (i.e., it is a "competitive" test) and is useful for singling out a few gene sets among many that are associated with a phenotype of interest. On the other hand, mROAST is a method that is focused only on genes within a set (it is "self-contained") and is more powerful for detecting subtle differences among phenotypes. The first workflow began with filtering the data to retain only genes with at least 10 read counts in greater than 5% of samples, followed by VST and removeBatchEffects. These data were then processed by gene set variation analysis (GSVA) to

produce enrichment scores for each sample and gene set [45]. Differential expression at the gene set level was assessed using a multivariate linear model and the empirical Bayes method in LIMMA.

Gene set analysis workflow 2: mROAST

The second gene set analysis workflow began with the same prevalence filtering followed by trimmed mean of M values (TMM) normalization and variance modeling at the observational level (VOOM) to generate precision weights [40]. We then performed gene set testing using the multiple rotation gene set test (mROAST, n rotations = 20,000, mean set statistic, mid-p values) [46]. RNA assay version was included as a covariate in mROAST.

For each data subset, models were run once for each of the imputed race and ethnicity categories with NH White as the reference group for four total tests, and once for each genetic ancestry proportion for five total tests, each on the appropriate set of pivot coordinates. To maximize robustness of findings, we required a Benjamini-Hochberg corrected $p < 0.05$ in both mROAST and GSVA to report a gene set as significantly enriched in a race and ethnicity imputed category, genetic ancestry proportion, or onset age group.

Consensus molecular subtypes

Consensus molecular subtypes (CMS) analysis was applied only to samples with colon or rectum as the sequenced tissue site [47, 48]. The CMScaller function assigned each sample a CMS, and a chi-squared test with post hoc inspection of standardized residuals was used to assess the relationship between CMS and imputed race and ethnicity categories. We further assessed this relationship stratified by age of onset category (EO vs. AO). For testing the association of CMS with genetic ancestry proportions, five separate multinomial logistic regressions were performed, each with the five CMS classes as dependent variables and genetic ancestry proportions (as pivot coordinate sets) as the independent variables. Finally, we repeated the multinomial logistic regression stratified by age of onset category.

Software

Somatic mutation analyses were performed with R version 4.1.3. RNA analyses were performed with R version 4.2.2. RNA-Seq data preparation and analysis steps are diagrammed in Additional file 1: Fig. S5.

Results

Associations between genetic ancestry and somatic mutations in MSS tumors

Among patients with MSS disease, we examined associations between genetic ancestry proportions and imputed race and ethnicity with protein-altering mutations in 79 genes (Additional file 1: Table S6, see "Methods" for selection criteria), somatic copy number alterations (SCNAs) in nine genes, and actionable mutations (present in OncoKB, cf. "Methods") in three genes (*BRAF*, *KRAS*, *PIK3CA*). Results for MSS tumors are summarized in Table 2 and Figs. 1 and 2.

Increased AFR ancestry was associated with higher odds of protein-altering mutations (Fig. 1A) in *APC* [odds ratio (OR) per doubling of ancestry proportion, 1.04; 95% confidence interval (CI), 1.02–1.06] and *KRAS* (OR, 1.04; 95% CI, 1.02–1.06), along with decreased odds of such mutations in *BRAF* (OR, 0.93; 95% CI, 0.90–0.97). EAS genetic ancestry was associated with decreased odds of protein-altering mutations in *KRAS* (OR, 0.98; 95% CI, 0.96–0.999). For actionable mutations (Fig. 1A), increased AFR genetic ancestry was associated with increased odds of *PIK3CA* mutations (OR, 1.04; 95% CI, 1.02–1.06) and decreased odds of *BRAF* mutations (OR, 0.90; 95% CI, 0.86–0.93). Increased EUR genetic ancestry was positively associated with actionable mutations in *BRAF* (OR, 1.09; 95% CI, 1.06–1.14). No genetic ancestry proportion associations were found with SCNAs. When including TMB (continuous) as a covariate, results were similar, with the exception of the EAS association with protein-coding mutations in *KRAS* and the AFR association with actionable mutations in *PIK3CA* no longer being statistically significant. Otherwise, all of the same associations were identified, with ORs and 95% confidence intervals changing by 0.01 or less (Additional file 1: Table S9).

In tests of imputed race and ethnicity categories, we found that NH Blacks had higher odds of protein-altering mutations (Fig. 1B) in *KRAS* compared to NH Whites (OR, 1.63; 95% CI, 1.37–1.94). The association was not significant for actionable mutations (Fig. 1B). NH Black and Hispanic/Latino patients had lower odds of actionable mutations of *BRAF* (OR, 0.61; 95% CI, 0.42–0.90 and OR, 0.29; 95% CI, 0.18–0.47, respectively) compared to NH White patients, while NH Blacks had higher odds of actionable mutations in *PIK3CA* (OR, 1.43; 95% CI, 1.18–1.75). When including TMB (continuous) as a covariate, we identified additional associations of protein-altering mutations in *BRAF* (OR, 0.44; 95% CI, 0.29–0.67) and *ERBB4* (OR, 0.36; 95% CI, 0.19–0.67) in NH Blacks compared to NH Whites. Otherwise, the same associations were identified (Additional file 1: Table S9).

Somatic mutation associations with interaction effects in MSS tumors

Two genes showed different imputed race and ethnicity category associations by either diagnosis age or primary site. Hispanic/Latino patients with AOCRC had higher

Table 2 Somatic mutation associations with ancestry proportions and imputed race categories in MSS patients. Mutation type: type of mutation tested. "Actionable" refers to protein-altering mutations that are classified as OncoKB Therapeutic Level of Evidence V2 designation of therapeutic level 1 or 2, or resistance level R1, irrespective of the solid cancer type. Gene: HGNC gene symbol of tested gene. *N* genes tested = number of genes of specified mutation type tested for association. *N* patients total: total number of patients included in models. *N* patients with mutation = number of patients included in models who have one or more of the mutation type in the gene. *p* LR (FDR): *p* value for likelihood ratio test, adjusted for the number of genes in the *N* genes tested column to control the false discovery rate. Ancestry or imputed race group: ancestry or imputed race group associated with the presence/absence of mutations in this gene in logistic regression test. OR (95% CI): odds ratio per doubling of genetic ancestry proportion (in the case of ancestry) or odds ratio compared to NH White category (in the case of imputed race group) and 95% confidence interval in the logistic regression test. *p* logistic = *p* value for the specific ancestry proportion or imputed race group in the logistic regression test, not adjusted for multiple tests

Mutation type	Gene	*N* genes tested	*N* patients total	*N* patients with mutation	*p* LR (FDR)	Ancestry or imputed race/ethnicity group	OR (95% CI)	*p* logistic
Protein-altering	*APC*	79	4871	3558	0.007	AFR	1.04 (1.02, 1.06)	1.6e−04
Protein-altering	*BRAF*	79	4871	365	0.047	AFR	0.93 (0.90, 0.97)	1.6e−04
Protein-altering	*KRAS*	79	4871	2343	0.004	AFR	1.04 (1.02, 1.06)	2.9e−06
						EAS	0.98 (0.96, 0.999)	0.039
Actionable	*BRAF*	3	7965	402	< 1e−08	AFR	0.90 (0.86, 0.93)	5.7e−08
						EUR	1.09 (1.06, 1.14)	5.4e−05
Actionable	*PIK3CA*	3	7965	855	0.016	AFR	1.04 (1.02, 1.06)	7.1e−04
Protein-altering	*KRAS*	79	4723	2276	3.2e−06	NH Black	1.63 (1.37, 1.94)	3.3e−08
Actionable	*BRAF*	3	7718	388	1.8e−08	Hispanic/Latino	0.61 (0.42, 0.90)	0.012
						NH Black	0.29 (0.18, 0.47)	3.9e−07
Actionable	*PIK3CA*	3	7718	831	0.003	NH Black	1.43 (1.18, 1.75)	3.4e−04

odds of *FLT3* SCNAs than NH White AOCRC patients (OR, 2.38; 95% CI, 1.53–3.72), while no association was present in those with EOCRC (OR, 0.46; 95% CI, 0.18–1.18; Fig. 3, Additional file 1: Table S8). Both NH Asian and NH Black patients with primary tumors in the colon showed decreased odds of actionable mutations in *BRAF* compared to NH Whites (OR, 0.12; 95% CI, 0.16–0.86 and OR, 0.10; 95% CI, 0.03–0.42, respectively), with no association seen for rectal tumors (OR, 1.84; 95% CI, 0.41–8.32 and OR, 2.12; 95% CI, 0.68–6.59, respectively) (cf. Figure 3, Additional file 1: Table S8).

Somatic mutation associations in MSI-high tumors

Approximately 5% of the cohort ($n = 445$) had MSI-high tumors, which ranged from 3.6 to 5.9% in patients with imputed NH Asian and Hispanic/Latino race and ethnicity, respectively (Table 1). The difference in proportion of MSI-high tumors by race and ethnicity category was not significant ($p = 0.062$). Among patients with AOCRC, prevalence of MSI-high tumors differed by imputed race and ethnicity ($p = 0.008$, Additional file 1: Table S5). Hispanic/Latino and NH White patients had the highest proportion of MSI-high tumors at 6.0% and 6.4%, respectively. In contrast, in EOCRC, patients with NH Black (5.4%), Hispanic/Latino (6.1%), and complex (6.9%) imputed race and ethnicity had higher rates of MSI-high tumors (Additional file 1: Table S6), though the differences were not statistically significant ($p = 0.074$). Among MSI-high patients, we tested the association of genetic ancestry proportions and imputed race and ethnicity with the presence of protein-altering mutations in 127 genes, SCNAs in two genes, and actionable mutations in two genes (Additional file 1: Table S6). No associations were found between genetic ancestry proportions and the presence of any mutations. NH Black MSI-high patients had higher odds of having protein-altering mutations in *KMT2C* compared to NH Whites (OR, 23.7; 95% CI, 3.1–181; Fig. 2A, Additional file 1: Tables S7–8), and NH Asian and Hispanic/Latino MSI-high patients were more likely to have *MLH1* SCNAs compared to NH Whites (OR, 13.9; 95% CI, 1.9–103 and OR, 11.4; 95% CI, 2.6–49.6, respectively (cf. Figure 2B, Additional file 1: Table S9)) were similar (Additional file 1: Table S10).

Somatic mutation association sensitivity tests

Sensitivity test results are given in Additional file 2: Tables S15–S24. In MSS tumors, associations with protein-altering somatic mutations in *KRAS*, *APC*, and *BRAF* were largely unchanged in sensitivity tests. LRT *p* values were > 0.05 in all tests that excluded patients with "not specified" cancer primary site, where statistical power was diminished due to a low number of patients

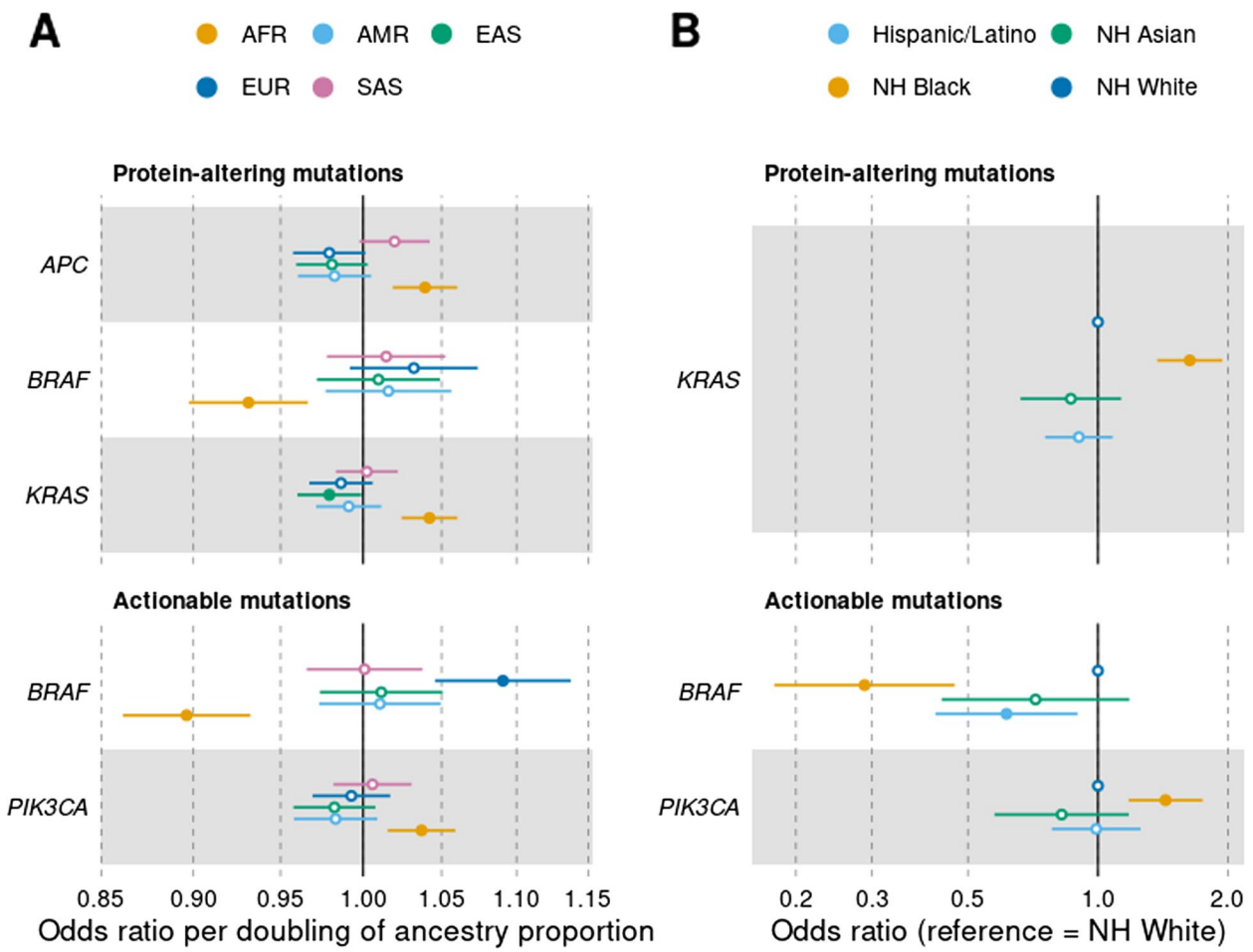

Fig. 1 Associations of somatic mutations with genetic ancestry proportions and imputed race and ethnicity categories in patients with MSS disease. **A** Associations with genetic ancestry proportions. AFR, Africa; AMR, the Americas; EAS, East Asia; EUR, Europe; SAS, South Asia. Odds ratios are with respect to a doubling of a specific genetic ancestry proportion and are adjusted for assay version, gender, age at sample collection, and the other four genetic ancestry proportions. **B** Associations with imputed race and ethnicity category. Odds ratios are with respect to the NH White race and ethnicity category and are adjusted for assay version, gender, and age at sample collection. Filled circles indicate a logistic regression $p < 0.05$, while open circles indicate $p \geq 0.05$

with specified cancer primary site. However, ORs from logistic regression were similar to the initial tests that adjusted only for age at collection, gender, and assay version (OR ranges: APC AFR 1.03–1.07, *BRAF* AFR 0.92–0.95, *KRAS* AFR 1.03–1.05, and *KRAS* EAS 0.97–0.98). Associations with actionable mutations in *BRAF* were all statistically significant, with ORs ranging from 0.88 to 0.93 for AFR and 1.08 to 1.12 for EUR. LRT p values for actionable mutations in *PIK3CA* were above 0.05 for tests restricted to patients with specified cancer primary site, and ORs for association with AFR genetic ancestry ranged from 1.02 to 1.05. In sensitivity tests of race and ethnicity categories among patients with MSS tumors, nearly all tests were statistically significant with a few exceptions in *PIK3CA* tests that excluded patients with unspecified cancer primary site, with ORs ranging from 1.54 to 1.85 for protein-altering *KRAS* mutations in NH Black patients, 0.25 to 0.37 for actionable *BRAF* mutations in NH Black patients, 0.57 to 0.73 for actionable BRAF mutations in Hispanic/Latino patients, and 1.26 to 1.72 for actionable *PIK3CA* mutations in NH Black patients.

LRT p values for sensitivity tests in patients with MSI-high tumors were all statistically significant, but logistic regression models that excluded patients with unspecified cancer primary site suffered from very low patient counts and perfect separation, resulting in extreme and unreliable ORs. The remaining ORs ranged from 14.3 to 26.1 for protein-altering *KMT2C* mutations in NH Black patients, 11.1 to 127.1 for *MLH1* SCNAs in NH Asian

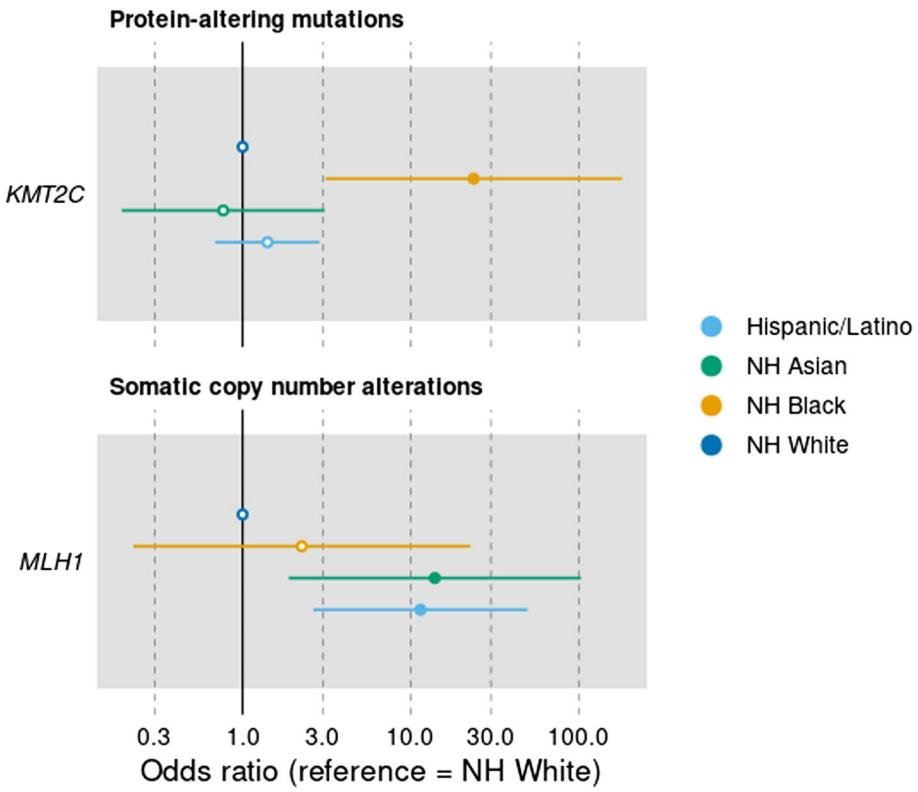

Fig. 2 Somatic mutation associations with imputed race and ethnicity categories in patients with MSI-high disease. Odds ratios are with respect to the NH White race and ethnicity category and are adjusted for assay version, gender, and age at sample collection. Filled circles indicate a logistic regression $p < 0.05$, while open circles indicate $p \geq 0.05$

patients, and 6.6 to 63.2 for *MLH1* SCNAs in Hispanic/Latino patients.

TMB in MSI-high tumors

Among MSI-high patients, there was no statistically significant difference of TMB by imputed race and ethnicity group ($p = 0.21$), or onset group ($p = 0.85$), nor was TMB significantly different among the subset of MSI-high patients with AOCRC ($p = 0.06$) or EOCRC ($p = 0.26$) (see Fig. 4A–D).

Variable selection for mRNA analyses

PCA plots of RNA counts demonstrated that after batch correction, tissue site was the primary driver of variation (Additional file 1: Fig. S2). Therefore, we restricted our analyses to liver, colon, and rectum samples (with liver assessed separately from colon/rectum) given small numbers in other metastatic sites (Table 1; Additional file 1: Table 11). Clinical stage was missing for 37% of patients with RNA-Seq results from the colon/rectum or liver, thus was not considered further. PCA plots were generated and labeled by MSI-status and tumor grade, and tumor tissue site for the colon/rectum subset. Given the small number of MSI-high patients (Additional file 1: Table S9) and differences in gene expression by MSI status (Additional file 1: Fig. S3), MSI-high patients were excluded from this analysis. There was notable clustering by missing tumor grade in both PCA and UMAP for the colon/rectum group with 21% of patients missing grade (Additional file 1: Fig. S4). Given the presence of strong clustering by grade, and grade likely missing not at random, a missing indicator approach was used. This method has been shown to produce an almost unbiased result while preserving power lost under complete case analysis [41]. Final variables included for multivariable analysis were tumor grade, gender, early versus average onset, colon vs. rectum tumor tissue site (not present for liver subgroup), and either imputed race and ethnicity categories or pivot coordinates for genetic ancestry proportions.

Associations between genetic ancestry and expression of gene sets

We next examined associations between genetic ancestry or imputed race and ethnicity category with expression of genes in the Hallmark and Biocarta C2 gene sets (342 total). In MSS colon/rectum samples ($n = 1830$), the imputed NH Black category was consistently

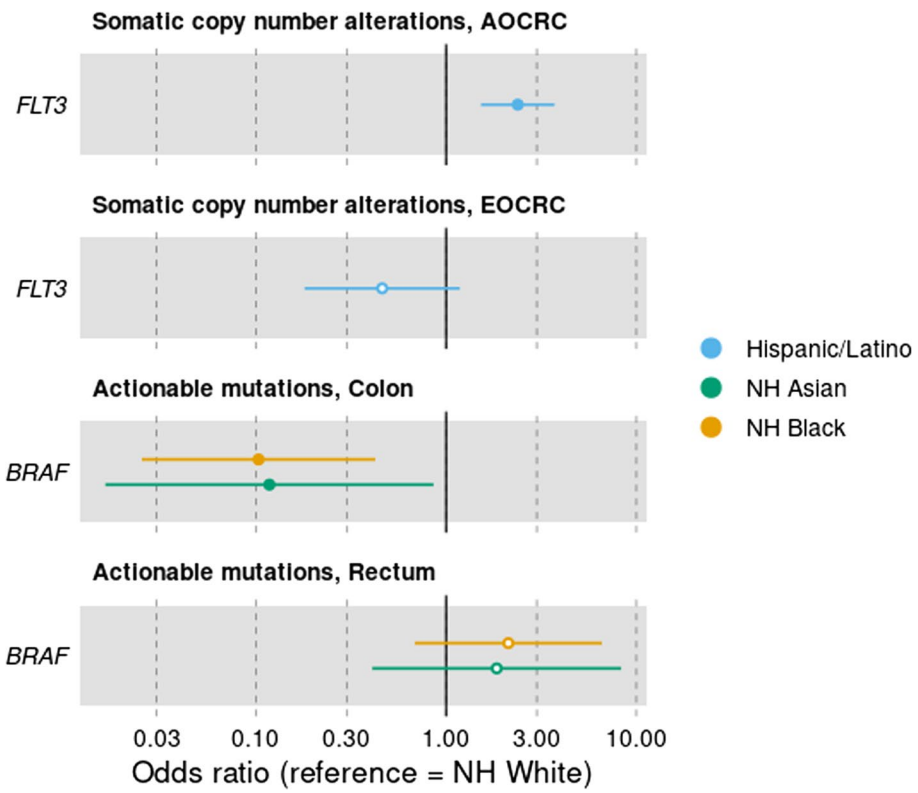

Fig. 3 Interaction effects by onset age group or by primary tumor site in somatic mutation associations with imputed race and ethnicity categories in patients with MSS disease. Odds ratios are with respect to the NH White race and ethnicity category and are adjusted for assay version, gender, and age at sample collection. Filled circles indicate a logistic regression $p < 0.05$, while open circles indicate $p \geq 0.05$

associated with underexpression compared to the NH White category in the following gene sets: Hallmark coagulation (mROAST $p = 0.021$, GSVA $p = 0.005$), BioCarta alternative complement (mROAST $p = 0.009$, GSVA $p = 0.005$), BioCarta RECK (mROAST $p = 0.026$, GSVA $p = 0.007$), and BioCarta Rhodopsin (mROAST $p = 0.026$, GSVA $p = 0.038$; Table 3). Highly differentially expressed genes in these gene sets included complement factor *C3*, tissue inhibitors of metalloproteinases *TIMP2* and *TIMP3*, matrix metallopeptidase 11 (*MMP11*), coagulation factor VIII (*F8*), cathepsin K (*CTSK*), and antithrombin III (*SERPINC1*) (Additional file 1: Table S12.1). Significant underexpression associated with increased AFR genetic ancestry in the above gene sets was found only by GSVA; we include the AFR results in Table 3 for comparison.

In MSS liver samples ($n = 778$), greater AFR genetic ancestry (but not the NH Black imputed category) was associated with underexpression in the BioCarta CREM gene set (Table 3 and Additional file 1: Table S12.2). There were no significant findings by age of onset group.

Associations between genetic ancestry and CRC consensus molecular subtypes (CMS)

Among 1957 patients where CMS were obtained with CMScaller, including both MSS and MSI-H patients, 252 were imputed non-Hispanic (NH) Black, 98 NH Asian, 287 Hispanic/Latino, 66 complex, and 1254 NH White (Additional file 1: Table 11). CMS was associated with race and ethnicity imputed categories ($p = 0.004$). Inspection of the standardized chi-square residuals revealed greater than expected NH Black CMS3 (66 observed vs. 46 expected, $p = 0.001$), less than expected NH Black CMS1 (18 vs. 30, $p = 0.011$), and greater than expected Hispanic/Latino indeterminate CMS (36 vs. 26, $p = 0.031$) (Fig. 5, Additional file 1: Table S13). When stratifying by age of onset group, the overall chi-square test of independence was no longer significant for EOCRC but remained significant for AOCRC, and inspection of standardized residuals revealed the association of indeterminate CMS and Hispanic/Latino imputed category was only present among EO, while the association of NH Black and CMS1 and CMS3 was only present among AO.

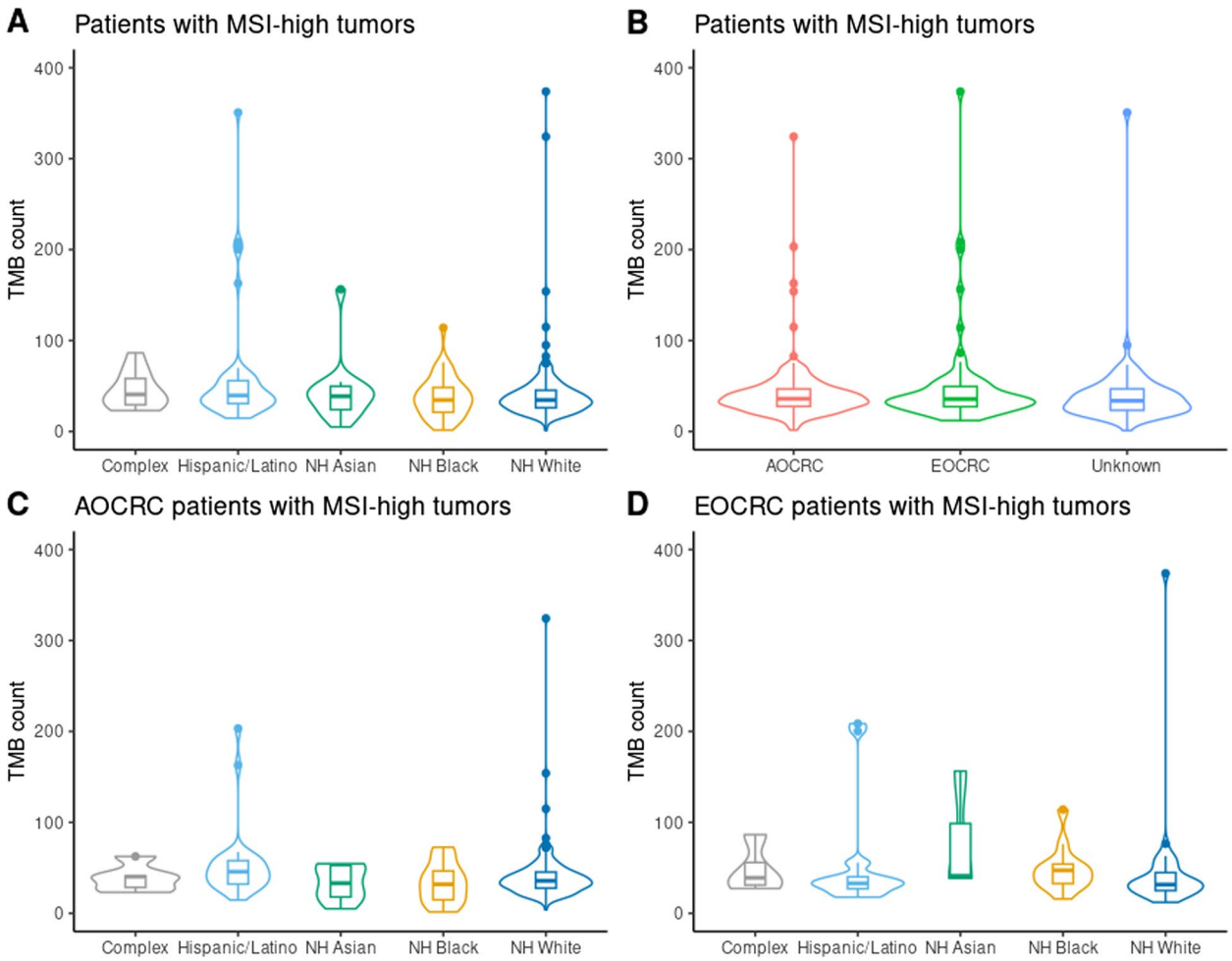

Fig. 4 Distribution of TMB count by imputed race and ethnicity group and age of onset for MSI-high patients. **A** TMB count by imputed race and ethnicity group. **B** TMB count by age of onset. **C** TMB count in AOCRC patients by race and ethnicity group. **D** TMB count in EOCRC patients by race and ethnicity group

In the analysis of genetic ancestry proportions, increased AFR genetic ancestry was significantly associated with CMS3 (OR, 1.056 per doubling of AFR proportion, 95% CI, 1.003–1.111) and indeterminate CMS (OR, 1.083; 95% CI, 1.021–1.149), while increased EUR genetic ancestry was associated with decreased odds of CMS3 (OR, 0.925; 95% CI, 0.874–0.979), all with CMS1 as the reference outcome (Additional file 1: Table S14). When stratifying by age of onset group, these associations were only statistically significant in the AOCRC group.

Discussion

In this molecular pathological epidemiology study, we utilized comprehensive tumor profiling in a large, diverse patient cohort derived from a clinico-genomic database, to identify differences in somatic mutation frequencies and gene expression by genetic ancestry in CRC. Unlike prior studies that have often relied on self-reported or observed race and ethnicity or rigid genetic ancestry categorizations, our approach directly employs ancestry proportions directly to identify associations, using statistical methods that control for correlations among ancestries. Further, we leverage genetic ancestry to impute race and ethnicity categories to address missingness in clinico-genomic databases.

Given the rising incidence of EOCRC, we first sought to assess differences in imputed race and ethnicity with tumor genetic profile by age of onset. However, in our data, no significant interactions were found except for *FLT3*, which had higher odds of SCNAs in Hispanic/Latino patients with AOCRC but not EOCRC.

For MSS CRC across all ages, NH Black patients and those with greater AFR genetic ancestry had lower odds of actionable variants in *BRAF* and higher odds of short protein-altering mutations in *KRAS*. In contrast, increased EAS ancestry was associated with decreased

Table 3 RNA gene set results. Gene sets are reported as significant only if the corrected *p* value from both mROAST and GSVA was < 0.05 and at least 50% of individual genes in the set were significantly differentially expressed as reported by mROAST. mROAST uses more conservative mid-*p* values during FDR correction. Included in this table alongside significant results is the result for the genetic ancestry proportion (or imputed group) that has the most overlap with the significant finding. All results reported in this table for the non-Hispanic Black imputed category are underexpression of the gene set in comparison to non-Hispanic White; all results for AFR represent decreased gene expression as the dominance of AFR compared to other ancestries increases

Gene set	Genetic ancestry	Tissue site	mROAST genes in set significantly underexpressed	mROAST FDR	GSVA FDR
Hallmark coagulation					
	NH Black	Colon/rectum	65/130	0.021	0.005
	AFR	Colon/rectum	60/130	0.083	0.038
BioCarta alternative complement					
	NH Black	Colon/rectum	5/8	0.009	0.005
	AFR	Colon/rectum	3/8	0.083	0.046
BioCarta RECK					
	NH Black	Colon/rectum	7/9	0.026	0.007
	AFR	Colon/rectum	7/9	0.105	0.042
BioCarta Rhodopsin					
	NH Black	Colon/rectum	5/6	0.026	0.011
	AFR	Colon/rectum	5/6	0.077	0.038
CREM					
	NH Black	Liver	3/5	0.528	0.079
	AFR	Liver	3/5	0.026	0.002

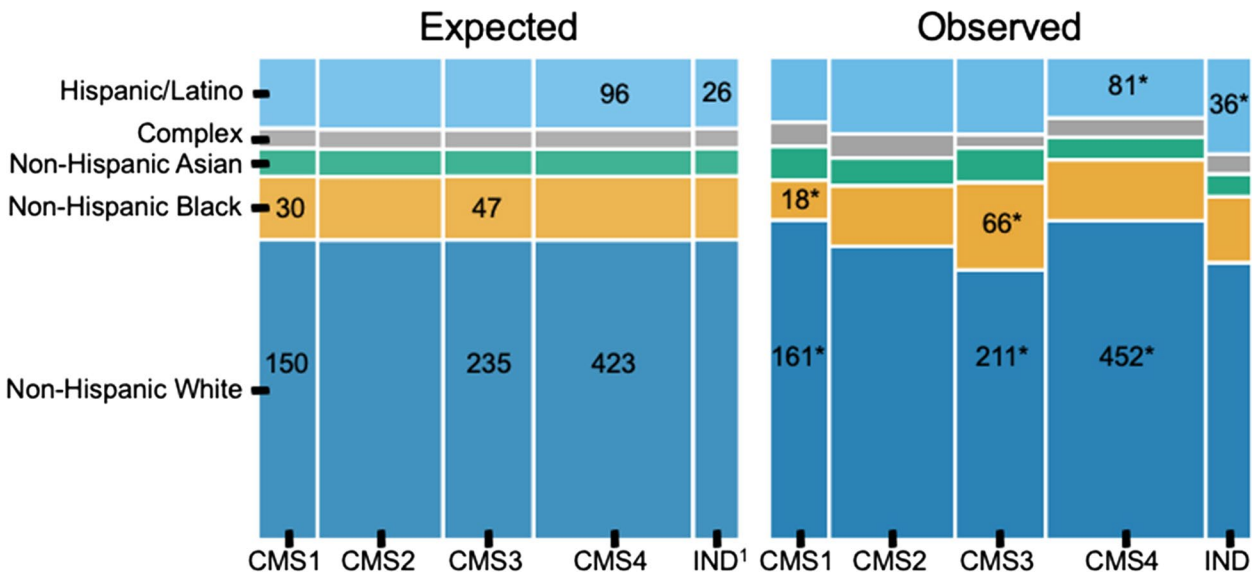

Fig. 5 Expected vs. observed proportions of patients by imputed racial/ethnic category and consensus molecular subtype. Panel area in the expected block is proportional to the null hypothesis of equal racial and ethnic distribution across CMS. Panel area in the observed block is proportional to the observed racial and ethnic distribution across CMS. Bar width for each CMS is proportional to the observed patient assignments. Findings with absolute standardized residuals > 1.96 indicated with * (chi-square test of independence) [1]Indeterminate CMS

odds of protein-altering *KRAS* mutations and greater odds of actionable *BRAF* mutations. *KRAS*-WT and left-sided tumors are approved for treatment with *EGFR* inhibitors such as cetuximab [49]. Our findings are consistent with previously published studies using self-reported race showing higher rates of *KRAS* mutations in NH Black patients [11, 12, 50], underscoring the importance of assessing targeted therapies in diverse populations. Standard of care for patients with metastatic cancer harboring mutations in *BRAF* V600E includes combination *BRAF* and *EGFR* inhibitors [51]. These mutations usually portend a poor prognosis with early development of metastatic disease and were less common in patients with imputed NH Black and Hispanic/Latino race and ethnicity and more common in patients with increased EUR ancestry [12].

Approximately 20–25% of CRC patients harbor activating mutations in *PIK3CA*, which activates the mTOR pathway [52]. Some studies have suggested that the presence of *PIK3CA* mutations could confer resistance to first-line chemotherapy, although the data are preliminary [53]. Inhibitors of PIK3CA have been approved for *PIK3CA* positive treatment resistant metastatic breast cancer [54]. As such, *PIK3CA* mutations could represent an opportunity for targeted therapy in CRC, particularly in combination with other drugs, since *PIK3CA* mutations are also associated with higher rates of mutations in genes in other key cancer pathways [52]. The higher rate of actionable *PIK3CA* mutations in MSS patients with greater AFR ancestry and imputed NH Black race suggests these combinations could preferentially benefit minority subgroups with CRC.

Thrombosis is one of the leading causes of death among cancer patients [55], and there is increased risk of both overall and cancer associated thrombosis among Black patients [56, 57]. In our study, the Hallmark coagulation gene set was significantly underexpressed in tumors from NH Black patients. Specifically, coagulation factors *F7* and *F11* and platelet tissue factor *TF* were underexpressed, while antithrombin III *SERPINC1* was overexpressed in NH Black patients compared to NH White. These findings do not support that changes in tumor coagulation gene expression pathways contribute to the elevated thrombosis risk observed in Black CRC patients.

Patients with imputed NH Black race and ethnicity or increased AFR ancestry had higher odds of CMS3 tumors. So-called metabolic tumors, CMS3 tumors display marked metabolic dysregulation with the majority harboring mutations in *KRAS* [47]. As such, this finding is concordant with the positive association of *KRAS* mutations and AFR ancestry found in our study. Hispanic/Latino patients were assigned to the indeterminate CMS category more often than expected. While the reason for this is unclear, one possibility is the underrepresentation of non-White patients in the datasets used to define CMS [32, 47]. As future trials and drug development efforts may stratify patients by CMS, it is important to ensure these categorizations accurately represent a diverse CRC patient population.

In spite of the advantages of our clinico-genomic database in terms of multimodality and greater diversity than controlled research and clinical studies, healthcare data are convenience samples with inherent ascertainment biases [20, 58]. Critical factors influencing the inclusion of patients—such as disease stage, race, ethnicity, insurance coverage, and socioeconomic status—are frequently unknown and unevenly distributed. This can lead to skewed data that may not accurately represent the broader population, potentially affecting the generalizability and validity of study findings. Limitations of our study include incomplete data on clinical stage, sidedness, age of diagnosis, and tumor grade, and unavailability of normal tissue sequencing for all patients. Our cohort has a larger representation of late-stage patients. Given that MSI-high is less common in late-stage metastatic patients and is associated with the CMS1 subtype, our results may be influenced by these differences. Missingness precluded the use of several variables as adjustment variables in our somatic mutation analyses. However, sensitivity tests indicated that our main results are unlikely to be totally explained by differences in tumor tissue site, cancer primary site, cancer stage, tumor grade, TMB, age at onset, cancer primary histology, or smoking status, though they may be partially explained by the latter two variables. It is important to note that our results should not be interpreted as indicating direct causal relationships between genetic ancestry or imputed race and ethnicity and molecular tumor profiles. Rather, these associations may be attributable to unmeasured genetic or environmental factors, or combinations thereof, that correlate with genetic ancestry proportions or imputed race and ethnicity. Additionally, data de-identification limits our ability to incorporate social determinants of health (SDOH) and other environmental factors that could influence more directly mutational profiles. This restriction prevents our molecular pathological epidemiology study from fully elucidating causality but opens the door for subsequent studies where these data are available. Furthermore, patients in our study do not represent an equal sampling of all patients across the United States, because our cohort consists of predominantly those with late-stage cancer whose physicians ordered the xT test. As such, our results may not generalize to all CRC patients, and some associations that exist in the full population may have been missed. Finally, we were not able to impute "American Indian

or Alaska Native" or "Native Hawaiian or Other Pacific Islander" categories due to limitations in the public reference allele frequencies and the small number of patients of such categories in our cohort (estimated < 1%). These patients may be misclassified as Hispanic/Latino or NH Asian, respectively (cf. Table 1); however, given the small number we do not expect these to significantly change our findings regarding imputed Hispanic/Latino or NH Asian categories.

In our study, we utilized imputed race and ethnicity due to the notable missingness of stated race and ethnicity data in healthcare and clinico-genomic data [15], where this information is not exclusively self-identified but also assigned by healthcare providers [14]. This approach significantly enhanced the statistical power to find associations while avoiding potential biases in data missingness [21, 59]. Our method leverages genetic ancestry for this imputation, and although genetic ancestry is not equivalent to race or ethnicity, a strong correlation between these two concepts has been observed among US populations [33]. We previously published an extensive analysis of the accuracy of our R/E imputation method and some variations, demonstrating that it outperforms other methods used in healthcare data [21]. We highlight that when performing race imputation, we adhered to established recommendations for ethical imputation—our adherence to these guidelines underscores our commitment to the responsible use of race imputation in promoting equity in healthcare [60].

Our cohort also includes a large fraction of patients for whom matched tumor-normal sequencing data is available, allowing better discrimination between germline variants and somatic mutations. Another strength of our study is the concurrent analysis of genomic somatic mutations with transcriptional profiles of the patient's tumors. Methodologically, by applying compositional analysis in our logistic regressions, we were able to minimize comparisons involving a single reference group (typically Whites) while controlling for correlations among genetic ancestries when they are reported as proportions that sum to one. Further, we used two distinct gene set analysis and RNAseq normalization methods to demonstrate consistency and strengthen our gene expression findings.

Conclusions

In summary, through analyzing a large, diverse CRC patient cohort, we found associations between genetic ancestry and prevalence of somatic mutations in CRC driver genes, gene expression levels in cancer related gene sets, and the distribution of consensus molecular subtypes that have not previously been reported in studies using race and ethnicity categories alone. Increased AFR genetic ancestry was associated with higher odds of *APC*, *KRAS*, and *PIK3CA* mutations and CMS3 tumors, as well as lower odds of *BRAF* mutations. Increased EAS genetic ancestry correlated with lower odds of mutations in *KRAS*. Furthermore, the increased odds of indeterminate CMS tumors in the imputed Hispanic/Latino category suggests that more diverse representation could reduce disparities in the applicability of disease subtype models. Additional work is needed to identify the specific genetic and environmental explanations of these associations. Our findings demonstrate the advantage of using genetic ancestry in studies of disparities in CRC and highlight the need to validate proposed therapies, biomarkers, and prognosis indicators in diverse patient populations.

Abbreviations

CRC	Colorectal cancer
EOCRC	Early onset CRC
AOCRC	Average onset CRC
CMS	CRC consensus molecular subtypes
AFR	African continental genetic ancestry
AIM	Ancestry-informative marker
AMR	American indigenous continental genetic ancestry
EAS	East Asian continental genetic ancestry
EUR	European continental genetic ancestry
SAS	South Asian continental genetic ancestry
NH	Non-Hispanic
MSI	Microsatellite instability
MSS	Microsatellite stable disease
SCNA	Somatic copy-number alterations
TMB	Tumor mutational burden
CI	Confidence interval
DE	Differential expression
GSVA	Gene set variation analysis
ILR	Isometric log ratio
LRT	Likelihood ratio test
OR	Odds ratio
PCA	Principal component analysis
TMM	Trimmed median of M values
VST	Variant stabilizing transform
RNA	Ribonucleic acid
RNAseq	RNA sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-024-01373-w.

Additional file 1: Supplementary methods, supplementary Tables S1–S14, and supplementary Figs. S1–S5

Additional file 2: Sensitivity analysis—Supplementary Tables S15–24

Acknowledgements

We would like to thank Yan Liu (Tempus), Carlos D. Bustamante, and Alex Ioannidis (Stanford University) for statistical and methodology discussions and advice. We also acknowledge Rafael Esteller, Nick Rigan, and Arvind Prasad from the Tempus Lens team, and Frank Nothaft, formerly from Tempus, for their superb assistance in procuring de-identified data and correcting data problems needed for this work. We thank Vanessa Nepomuceno from the Tempus Publications team for copyediting the manuscript.

Authors' contributions
Brooke Rhead: methodology, data analysis, visualization, writing—review, editing. David Hein: methodology, data analysis, visualization, writing—review, editing. Yannick Pouliot: data procurement, curation, analysis, methodology, writing—review, editing. Justin Guinney: methodology, writing—review and editing. Francisco De La Vega: conceptualization, resources, supervision, writing—review, editing. Nina Sanford: conceptualization, resources, supervision, writing—original draft, writing—review, editing. All authors reviewed and suggested edits for the final version of the manuscript. The authors read and approved the final manuscript.

Funding
This study was funded by Tempus AI, Inc.

Availability of data and materials
De-identified, individual-level data used in this research was collected in a real-world healthcare setting and is subject to controlled access for privacy and contractual reasons. The ethics committee and/or informed consent does not allow for public availability. Derived data supporting the conclusions of this article are included within the article and its additional files. Tempus may make access to further data pending a signed data use agreement. Requests for access should be sent to publication.inquiry@tempus.com. For further information, visit https://www.tempus.com/life-sciences/data-collaborations.

Declarations

Ethics approval and consent to participate
All analyses were performed using de-identified data and therefore were not considered human subjects research. As such, the need for Institutional Review Board (IRB) approval was exempted by the IRB of Advarra, Inc., under protocol number Pro00042950, on April 15, 2020. While this work is not classified as human subjects research, it conforms to the ethical principles of the Helsinki Declaration.

Consent for publication
Not applicable.

Competing interests
B.R., Y.P., J.G., and F.M.D.L.V. are or were employees and have received restricted stock from Tempus AI, Inc. The remaining authors declare that they have no competing interests. This research was funded by Tempus AI, Inc.

Author details
[1]Tempus AI, 600 West Chicago Avenue, Suite 510, Chicago, IL 60654, USA. [2]Department of Radiation Oncology, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390, USA. [3]Department of Biomedical Data Science, Stanford University School of Medicine, 1265 Welch Road, Stanford, CA 94305, USA.

Received: 14 August 2023 Accepted: 5 August 2024
Published online: 13 August 2024

References
1. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin. 2020;70(3):145–64. https://doi.org/10.3322/caac.21601.
2. Murphy CC, Wallace K, Sandler RS, Baron JA. Racial disparities in incidence of young-onset colorectal cancer and patient survival. Gastroenterology. 2019;156(4):958–65. https://doi.org/10.1053/j.gastro.2018.11.060.
3. Sinicrope FA. Increasing incidence of early-onset colorectal cancer. Longo DL, ed. N Engl J Med. 2022;386(16):1547–58. https://doi.org/10.1056/NEJMra2200869.
4. Mork ME, You YN, Ying J, et al. High prevalence of hereditary cancer syndromes in adolescents and young adults with colorectal cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2015;33(31):3544–9. https://doi.org/10.1200/JCO.2015.61.4503.
5. Cercek A, Chatila WK, Yaeger R, et al. A comprehensive comparison of early-onset and average-onset colorectal cancers. JNCI J Natl Cancer Inst. 2021;113(12):1683–92. https://doi.org/10.1093/jnci/djab124.
6. Lieu CH, Golemis EA, Serebriiskii IG, et al. Comprehensive genomic landscapes in early and later onset colorectal cancer. Clin Cancer Res. 2019;25(19):5852–8. https://doi.org/10.1158/1078-0432.CCR-19-0899.
7. McLeod MR, Galoosian A, May FP. Racial and ethnic disparities in colorectal cancer screening and outcomes. Hematol Oncol Clin North Am. 2022;36(3):415–28. https://doi.org/10.1016/j.hoc.2022.02.003.
8. Carethers JM. Clinical and genetic factors to inform reducing colorectal cancer disparitites in African Americans. Front Oncol. 2018;8:531. https://doi.org/10.3389/fonc.2018.00531.
9. Lai Y, Wang C, Civan JM, et al. Effects of cancer stage and treatment differences on racial disparities in survival from colon cancer: a United States population-based study. Gastroenterology. 2016;150(5):1135–46. https://doi.org/10.1053/j.gastro.2016.01.030.
10. Hein DM, Deng W, Bleile M, et al. Racial and ethnic differences in genomic profiling of early onset colorectal cancer. JNCI J Natl Cancer Inst. 2022;114(5):775–8. https://doi.org/10.1093/jnci/djac014.
11. Myer PA, Lee JK, Madison RW, et al. The genomics of colorectal cancer in populations with African and European ancestry. Cancer Discov. 2022;12(5):1282–93. https://doi.org/10.1158/2159-8290.CD-21-0813.
12. Yoon HH, Shi Q, Alberts SR, et al. Racial differences in *BRAF/KRAS* mutation rates and survival in stage III colon cancer patients. J Natl Cancer Inst. 2015;107(10):djv186. https://doi.org/10.1093/jnci/djv186.
13. Nead KT, Hinkston CL, Wehner MR. Cautions when using race and ethnicity in administrative claims data sets. JAMA Health Forum. 2022;3(7):e221812. https://doi.org/10.1001/jamahealthforum.2022.1812.
14. White K, Lawrence JA, Tchangalova N, Huang SJ, Cummings JL. Socially-assigned race and health: a scoping review with global implications for population health equity. Int J Equity Health. 2020;19(1):25. https://doi.org/10.1186/s12939-020-1137-5.
15. Studna A. The rise of RWD in clinical research. Appl Clin Trials. 2023;32(5). https://www.appliedclinicaltrialsonline.com/view/executive-roundtable-the-rise-of-rwd-in-clinical-research. Accessed 16 Jul 2023.
16. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Hum Genomics. 2015;9(1):1. https://doi.org/10.1186/s40246-014-0023-x.
17 Borrell LN, Elhawary JR, Fuentes-Afflick E, et al. Race and genetic ancestry in medicine — a time for reckoning with racism. Malina D, ed. N Engl J Med. 2021;384(5):474–80. https://doi.org/10.1056/NEJMms2029562.
18. Revisions to the standards for the classification of federal data on race and ethnicity. Published online October 30, 1997. https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf. Accessed 26 May 2022.
19. Beaubier N, Bontrager M, Huether R, et al. Integrated genomic profiling expands clinical options for patients with cancer. Nat Biotechnol. 2019;37(11):1351–60. https://doi.org/10.1038/s41587-019-0259-z.
20. Spratt DE, Chan T, Waldron L, et al. Racial/ethnic disparities in genomic sequencing. JAMA Oncol. 2016;2(8):1070. https://doi.org/10.1001/jamaoncol.2016.1854.
21. Rhead B, Haffener PE, Pouliot Y, De La Vega FM. Imputation of race and ethnicity categories using genetic ancestry from real-world genomic testing data. In: Biocomputing 2024. WORLD SCIENTIFIC; 2023. p. 433–445. https://doi.org/10.1142/9789811286421_0033.
22. Beaubier N, Tell R, Lau D, et al. Clinical validation of the tempus xT next-generation targeted oncology sequencing assay. Oncotarget. 2019;10(24):2384–96. https://doi.org/10.18632/oncotarget.26797.
23. Sanchez-Vega F, Mina M, Armenia J, et al. Oncogenic signaling pathways in The Cancer Genome Atlas. Cell. 2018;173(2):321–337.e10. https://doi.org/10.1016/j.cell.2018.03.035.
24. Martínez-Jiménez F, Muiños F, Sentís I, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555–72. https://doi.org/10.1038/s41568-020-0290-x.
25. Chakravarty D, Gao J, Phillips S, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;1:1–16. https://doi.org/10.1200/PO.17.00011.
26. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64. https://doi.org/10.1101/gr.094052.109.
27. Miyashita M, Bell JSK, Wenric S, et al. Molecular profiling of a real-world breast cancer cohort with genetically inferred ancestries reveals

actionable tumor biology differences between European ancestry and African ancestry patient populations. Breast Cancer Res. 2023;25(1):58. https://doi.org/10.1186/s13058-023-01627-2.
28. The 1000 Genomes Project Consortium, Corresponding authors, Auton A, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. https://doi.org/10.1038/nature15393.
29 Bergström A, McCarthy SA, Hui R, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367(6484):eaay5012. https://doi.org/10.1126/science.aay5012.
30. Mallick S, Li H, Lipson M, et al. The simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201–6. https://doi.org/10.1038/nature18964.
31. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93(2):278–88. https://doi.org/10.1016/j.ajhg.2013.06.020.
32. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Aaltonen LA, Abascal F, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93. https://doi.org/10.1038/s41586-020-1969-6.
33. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet. 2015;96(1):37–53. https://doi.org/10.1016/j.ajhg.2014.11.010.
34. Templ M, Hron K, Filzmoser P. robCompositions: an R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn V, Buccianti A, eds. Compositional data analysis. 1st ed. Wiley; 2011. p. 341–355. https://doi.org/10.1002/9781119976462.ch25
35. Marabelle A, Fakih M, Lopez J, et al. Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. Lancet Oncol. 2020;21(10):1353–65. https://doi.org/10.1016/S1470-2045(20)30445-9.
36. Cercek A, Lumish M, Sinopoli J, et al. PD-1 blockade in mismatch repair–deficient, locally advanced rectal cancer. N Engl J Med. 2022;386(25):2363–76. https://doi.org/10.1056/NEJMoa2201445.
37. Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible summary tables with the gtsummary package. R J. 2021;13(1):570. https://doi.org/10.32614/RJ-2021-053.
38. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7. https://doi.org/10.1038/nbt.3519.
39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
40. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29. https://doi.org/10.1186/gb-2014-15-2-r29.
41. Zhuchkova S, Rotmistrov A. How to choose an approach to handling missing categorical data: (un)expected findings from a simulated statistical experiment. Qual Quant. 2022;56(1):1–22. https://doi.org/10.1007/s11135-021-01114-w.
42 Nishimura D. BioCarta. Biotech Softw Internet Rep. 2001;2(3):117–20. https://doi.org/10.1089/152791601750294344.
43. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Syst. 2015;1(6):417–25. https://doi.org/10.1016/j.cels.2015.12.004.
44. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
45. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013;14(1):7. https://doi.org/10.1186/1471-2105-14-7.
46. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinforma Oxf Engl. 2010;26(17):2176–82. https://doi.org/10.1093/bioinformatics/btq401.
47. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21(11):1350–6. https://doi.org/10.1038/nm.3967.
48. Eide PW, Bruun J, Lothe RA, Sveen A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci Rep. 2017;7(1):16618. https://doi.org/10.1038/s41598-017-16747-x.
49. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. *K-ras* mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med. 2008;359(17):1757–65. https://doi.org/10.1056/NEJMoa0804385.
50. Staudacher JJ, Yazici C, Bul V, et al. Increased frequency of KRAS mutations in African Americans compared with Caucasians in sporadic colorectal cancer. Clin Transl Gastroenterol. 2017;8(10): e124. https://doi.org/10.1038/ctg.2017.48.
51. Tabernero J, Grothey A, Van Cutsem E, et al. Encorafenib plus cetuximab as a new standard of care for previously treated *BRAF* V600E–mutant metastatic colorectal cancer: updated survival results and subgroup analyses from the BEACON study. J Clin Oncol. 2021;39(4):273–84. https://doi.org/10.1200/JCO.20.02088.
52. Voutsadakis IA. The landscape of PIK3CA mutations in colorectal cancer. Clin Colorectal Cancer. 2021;20(3):201–15. https://doi.org/10.1016/j.clcc.2021.02.003.
53. Wang Q, Shi YL, Zhou K, et al. PIK3CA mutations confer resistance to first-line chemotherapy in colorectal cancer. Cell Death Dis. 2018;9(7):739. https://doi.org/10.1038/s41419-018-0776-6.
54. Narayan P, Prowell TM, Gao JJ, et al. FDA approval summary: alpelisib plus fulvestrant for patients with HR-positive, HER2-negative, PIK3CA-mutated, advanced or metastatic breast cancer. Clin Cancer Res. 2021;27(7):1842–9. https://doi.org/10.1158/1078-0432.CCR-20-3652.
55. Lewis-Lloyd CA, Pettitt EM, Adiamah A, Crooks CJ, Humes DJ. Risk of postoperative venous thromboembolism after surgery for colorectal malignancy: a systematic review and meta-analysis. Dis Colon Rectum. 2021;64(4):484–96. https://doi.org/10.1097/DCR.0000000000001946.
56. Datta T, Brunson A, Mahajan A, Keegan T, Wun T. Racial disparities in cancer-associated thrombosis. Blood Adv. 2022;6(10):3167–77. https://doi.org/10.1182/bloodadvances.2021006209.
57. Key NS, Reiner AP. Genetic basis of ethnic disparities in VTE risk. Blood. 2016;127(15):1844–5. https://doi.org/10.1182/blood-2016-03-701698.
58. Verkerk K, Voest EE. Generating and using real-world data: a worthwhile uphill battle. Cell. 2024;187(7):1636–50. https://doi.org/10.1016/j.cell.2024.02.012.
59. Srivastav A, Robinson-Ector K, Kipp C, Strompolis M, White K. Who declines to respond to the reactions to race module?: findings from the South Carolina Behavioral Risk Factor Surveillance System, 2016–2017. BMC Public Health. 2021;21(1):1703. https://doi.org/10.1186/s12889-021-11748-y.
60. Brown KS, Ford L, Ashley S, Stern A, Ajjit N. Ethics and empathy in using imputation to disaggregate data for racial equity: recommendations and standards guide. Washington, DC: Urban Institute; 2021.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.