Equity in cancer genomics in the UK: a cross-sectional analysis of a national cancer cohort

T Nguyen, Sam Tallman*, Yoonsu Cho, Alona Sosinsky, John Ambrose, Steve Thorn, Maxine Mackintosh, Matthew A Brown, Loukas Moutsianas, Matt J Silver†, Karoline Kuchenbaecker†*

Summary

Background Most research on genetic screening and precision oncology is based on individuals of European ancestry. We applied the National Health Service (NHS) England's cancer variant prioritisation workflow to evaluate the performance of these approaches in ethinically and ancestrally diverse populations. The second aim of the study was to assess the representativeness of the 100 000 Genomes Project cancer cohort of the population of England.

Methods In this cross-sectional analysis, whole-genome sequencing data from patients with cancer recruited into the 100 000 Genomes Project between February 2015 to December 2018 were analysed. Clinical information, including tumour stage and grade, was gathered from the NHS England National Cancer Registration and Analysis Service. Patients with cancer types with fewer than five individuals, haematological cancers, childhood cancers, unknown primary carcinomas, patients with indeterminate sex, and patients missing somatic mutations in genes were excluded. To assess ethnicity representation in the 100 000 Genomes Project, we calculated the recruitment ratios for self-reported ethnicities for patients with cancer recruited to the 100 000 Genomes Project and patients with cancer in England. We also analysed differences in classification rates for potentially pathogenic variants to assess ancestry-related differences in germline and somatic mutations of different ancestry groups.

Findings 14 775 patients with cancer were recruited between February, 2015, and December, 2018, into the 100 000 Genomes Project. There was no evidence of under-representation of diverse ethnic groups in the 100 000 Genomes Project when compared with the national statistics. The recruitment rate ratio for breast cancer was 2·2 (95% CI 1·6–3·0) for Black versus White women in the 100 000 Genomes Project compared with 0·81 (0·79–0·83) for Black versus White women in the national data (fold-change in rate ratios 2·7; 95% CI 2·0–3·7, p<0·0001), suggesting higher representation of Black women in the 100 000 Genomes Project than expected given the ethnicity-specific incidence rates in England. Compared with national rates, the 100 000 Genomes Project also had higher recruitment rates of Black versus White men with prostate cancer (fold-change in rate ratios 3·7; 1·8–7·5, p=0·0004), Black versus White men with bladder cancer (fold change in rate ratios 6·1; 2·0–18·8, p=0·0016), and Asian versus White women with breast cancer (fold change in rate ratios 1·4; 1·2–1·7, p=0·0008). Ancestry had a significant association with the likelihood of carrying a variant classified as a potentially pathogenic (likelihood ratio test p=0·0011). Potentially pathogenic variants were identified in 23 (4·6%) of 500 South Asian (adjusted model odds ratio [OR] 1·88, 95% CI 1·21–2·93, p=0·0052) and 24 (5·3%) of 453 African ancestry patients (OR 2·24, 1·44–3·48, p=0·0003) compared with 263 (2·2%) of 11 955 in European-ancestry patients. However, we found that fewer tumour mutations in actionable genes were identified for patients of non-European ancestry compared with patients of European ancestry when adjusting for sex and cancer type (likelihood ratio test p<0·0001).

Interpretation There was an excess of germline variants classified as potentially pathogenic variants in patients with non-European ancestry, which might impede the diagnostic process. Improved variant prioritisation workflows and more research in diverse groups are needed to ensure equitable implementation of genomics in cancer care.

Funding The UK Department of Health and Social Care and the EU's Horizon 2020 Research and Innovation Programme.

Lancet Oncol 2025; 26: 971–80

Published Online June 10, 2025 https://doi.org/10.1016/S1470-2045(25)00199-8

See Comment page 827

*Contributed equally as first authors

†Contributed equally as last authors

Genomics England, London, UK (T Nguyen MSc, S Tallman PhD, Y Cho, PhD, A Sosinsky PhD, J Ambrose PhD, M Mackintosh PhD, Prof M A Brown PhD, L Moutsianas PhD, M J Silver PhD, Prof K Kuchenbaecker PhD); **Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK** (Y Cho); **Department of Oncology, University of Oxford, Oxford, UK** (S Thorn PhD); **Alan Turing Institute, London, UK** (M Mackintosh); **Department of Non-Communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, UK** (M J Silver); **Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, Banjul, The Gambia** (M J Silver); **Division of Psychiatry, University College London, London, UK** (Prof K Kuchenbaecker); **UCL Genetics Institute, University College London, London, UK** (Prof K Kuchenbaecker)

Correspondence to: Prof Karoline Kuchenbaecker, Division of Psychiatry, University College London, London WC1E 6BT, UK **k.kuchenbaecker@ucl.ac.uk**

Introduction

Identification of individuals with cancer susceptibility variants through genome sequencing facilitates prevention and earlier disease diagnosis. Characterising somatic mutations in tumour tissue has an important role in targeted treatment. To our knowledge, England was the first country to offer whole-genome sequencing as part of routine cancer care within a national health-care system, through the Genomic Medicine Service of National Health Service (NHS) England.[1] The foundation for this service was laid by the 100 000 Genomes Project,[2] which recruited more than 15 000 patients to the programme in 2015–18.[3] Several other countries are now following suit to increase the use of whole-genome sequencing in oncology.[4]

Previous studies have identified differences between ethnicities in terms of cancer incidences, diagnoses, and

Research in context

Evidence before this study

We reviewed the existing literature on cancer genomics and disparities related to ancestry using PubMed. The search covered articles published from database inception to April 1, 2024, and used keywords: "cancer" and "cancer genomics", "ancestry disparities", "variant prioritization", "precision oncology," and "genetic diversity". No other search restrictions were applied. We included studies that evaluated genetic variant classification or cancer care in diverse populations and excluded those that did not have ancestry-specific analysis or genomic prioritisation data. Evidence consistently showed poor or no representation of individuals of diverse ancestry in cancer genomics research, limiting our understanding of the performance of precision oncology across ancestries. USA-based studies found that screening using gene panels yields more variants of uncertain significance for Black and Asian individuals and lower rates of pathogenic variants compared with White individuals. However, a study on breast cancer published in 2021 did not find any statistically significant differences. To our knowledge, outside the USA, no large study has explored germline and somatic variant classification discrepancies in cancer genomic data.

Added value of this study

This study offers a unique analysis of cancer genomics disparities by ancestry using data from the 100 000 Genomes Project, which includes over 14 000 patients from diverse ethnic and ancestral backgrounds. Adapting the cancer variant prioritisation workflow used by the English National Health Service's Genomic Medicine Service, we observed significant disparities in variant prioritisation across ancestries. Patients of non-European ancestry were more likely to have germline variants classified as potentially pathogenic. Additionally, we observed fewer somatic mutations in actionable genes in non-European ancestry groups than in the European ancestry group.

Implications of all the available evidence

These findings call for refinements in the variant prioritisation workflow and variant annotation to improve genetic screening for individuals with non-European ancestries, combined with the need for enhanced genomic reference datasets that better represent genetic diversity. Findings also raise concerns over unequal clinical benefits of tumour sequencing and point to an urgent need for dedicated research to develop targeted therapies based on diverse patient populations.

prognoses.[5–8] Genetic screening and precision oncology must provide equal benefits for people from diverse ethnic backgrounds to avoid exacerbating existing health inequalities. A US-based study found that genetic screening using gene panels yields variants of uncertain significance about twice as often for individuals who are Black or Asian compared with individuals who are White.[9] Patients who are not White also had slightly lower rates of variants that were likely to be classified as germline pathogenic.[10] However, a later study did not find a statistically significant difference in pathogenic variants between Black and White women with breast cancer.[11]

See Online for appendix

The influence of genetic ancestry on genetic screening and precision oncology has never been investigated for a national cancer sequencing programme. Furthermore, the causes of any genetic ancestry-related differences in cancer incidences, diagnoses, and prognoses remain largely unknown, so that routes towards reducing disparities are currently unclear.

In this Article, we aim to examine how representative the 100 000 Genomes Project Cancer Programme is in terms of self-reported ethnicity, compared with national statistics. We also assess whether there are ancestry-related differences in the identification of clinically relevant germline variants in high-risk genes, such as *BRCA1* or *BRCA2*, as well as in the identification of somatic cancer variants and investigate how these relate to demographics, tumour characteristics, population genetic differences, and features of the bioinformatics pipeline used for variant prioritisation.

For the **100 000 Genomes Project study protocols** please see https://www.genomicsengland.co.uk/initiatives/100000-genomes-project/documentation

Methods

Study design

In this cross-sectional analysis, we analysed whole-genome sequencing data from patients with cancer recruited into the 100 000 Genomes Project[3] between February, 2015, and December, 2018 (appendix pp 3–4). Our analysis was limited to patients in the version 15 release of the National Genomic Research Library, a database of whole-genome sequence and linked health data from consenting NHS patients. Each patient had one germline sample sequenced at 30× coverage and one or more tumour samples sequenced at 100× coverage. The whole-genome sequence results were returned between 2016 and 2019. Cancer types were assigned according to the location of the primary tumour.

Ethical approval for the 100 000 Genomes Project was obtained from the East of England–Cambridge South Research Ethics Committee (reference 14/EE/1112, integrated research application system identification 166046). Documents related to the 100 000 Genomes Project study protocols are available online. Patient representatives were not involved in shaping the current study.

Patients

Basic patient and sample data were collected at the time of DNA sample submission to the 100 000 Genomes Project. Secondary clinical information, including cancer stage and tumour grade, was gathered from the NHS England National Cancer Registration and Analysis Service.[3] Patients with cancer types with fewer than five individuals,

haematological cancers, childhood cancers, unknown primary carcinomas, patients with indeterminate sex, and patients missing somatic mutations in genes were excluded. Patients of admixed Latin American ancestry were also excluded due to low sample numbers. 100 000 Genomes Project participant ethnicity was self-reported at registration using UK Office of National Statistics categories.[12] Written informed consent was obtained from participating patients, including for genetic testing and reporting test results back to their clinical team before discussions with the patient.

Procedures

Whole-genome sequence data of each patient was used to assigned them to genetically inferred ancestry groups based on their similarity to so-called super populations—European, African, South Asian, and Admixed American—in the 100 000 Genomes Project, as described elsewhere (appendix p 5).[13] Any patient who did not meet the threshold for similarity to a single reference population (ie, ≥0·8 probability of belonging to a specific super population class) was labelled as unassigned. In this Article, we occasionally use terms for brevity, such as "patients with European ancestry" to refer to patients assigned to the European ancestry group, while noting that no individuals derive from a single ancestry in any meaningful sense. When referring to ancestry in this Article, genetically inferred ancestry is always implied.

To align with clinical practices in England, we adopted the variant prioritisation workflow used for patients with cancer within NHS England's Genomic Medicine Service. Germline and somatic variants were identified and classified by the 100 000 Genomes Project cancer bioinformatics pipeline (version 1.6–1.11).[14] Functional, allele frequency, and variant database annotations were obtained from the National Genomic Research Library.[15]

Germline variants were classified as likely pathogenic or pathogenic variants (ie, Tier 1 in the Genomics England pipeline) if they were found in genes linked to the patient's cancer type according to PanelApp, a gene panel database for health disorders,[16] and were either predicted protein truncating variants, for which the mechanism of pathogenicity is loss of function (excluding variants annotated as benign or likely benign in Clinvar[17]), or were listed in ClinVar as pathogenic or likely pathogenic (with a rating of at least two stars; appendix pp 19–20). We refer to these as germline pathogenic variants.

Candidate variants (ie, Tier 3 in the Genomics England workflow) are a lower priority classification, across a wider range of consequence types, including missense variants, where the frequency in an internal Genomics England dataset of over 6000 unrelated individuals is less than 0·05% (rare variant threshold) for dominant-acting genes and less than 2% (common variant threshold) for recessive genes. Candidate variants included genes within a broader cancer susceptibility panel as well as familial cancer syndromes. Variants listed in ClinVar as benign or likely benign with a rating of at least two stars were excluded from pathogenic variants and candidate variants.

Somatic mutations were classified as actionable, or domain 1, if they were protein altering and located in genes affecting diagnosis, prognosis, or treatment for the patient's cancer type or if they led to eligibility for a clinical trial. Cancer-related somatic domain 2 mutations were protein-altering mutations located in genes implicated in any cancer type. Domain 3 mutations were protein-altering mutations found in any protein-coding genes. (appendix p 21). We only included data for one tumour sample per patient in the analysis.

Statistical analysis

We compared recruitment rate ratios of Black to White and Asian to White ethnicities with those reported by Delon and colleagues.[5] In their analysis, Delon and colleagues reported age-standardised incidence rates for ethnic groups in England in 2013–17, and ratios of the age-standardised incidence rate for each non-White ethnicity versus White ethnicity. If recruitment of patients to the 100 000 Genomes Project for cancer was representative, then the ratios of the age-standardised recruitment rates should be similar to the ratios of the age-standardised incidence rates for England. Therefore, we calculated Black to White and Asian to White recruitment rate ratios for 100 000 Genomes Project participants for each cancer type using Delon and colleagues' method and compared the resultant recruitment rate ratios with the previously reported ratios. Only cancer types with at least 100 patients were compared. We compared these rate ratios using z-tests and applied a Bonferroni correction to resulting p values.

The association of genetic ancestry with the probability of finding at least one pathogenic variant was modelled by logistic regression with cancer type as a random effect. Negative binomial regression was used to evaluate the association between ancestry and the total number of candidate germline variants because many patients carried one or more of these variants. We also used negative binomial regression for the association analyses between ancestry and the number of non-synonymous somatic variants in actionable genes as well as ancestry and number of somatic mutations of all domains.

Likelihood ratio tests were used to test for an overall effect of ancestry on outcomes. For significant likelihood ratio tests ($p<0·05$), we assessed differences between individual groups using the regression coefficients for the categorical ancestry group predictors.

We explored whether ancestry differences in germline variants are linked to other variables, including age at registration, ratio of heterozygous to homozygous variants, and total number of germline variants, by adding them as covariates to the model. For somatic mutations, the variables we considered were age at registration, tumour mutation burden (TMB), tumour grade, and cancer type. TMB was defined as the count of somatic

For the **panel of genes for each cancer type** see https://panelapp.genomicsengland.co.uk/panels/

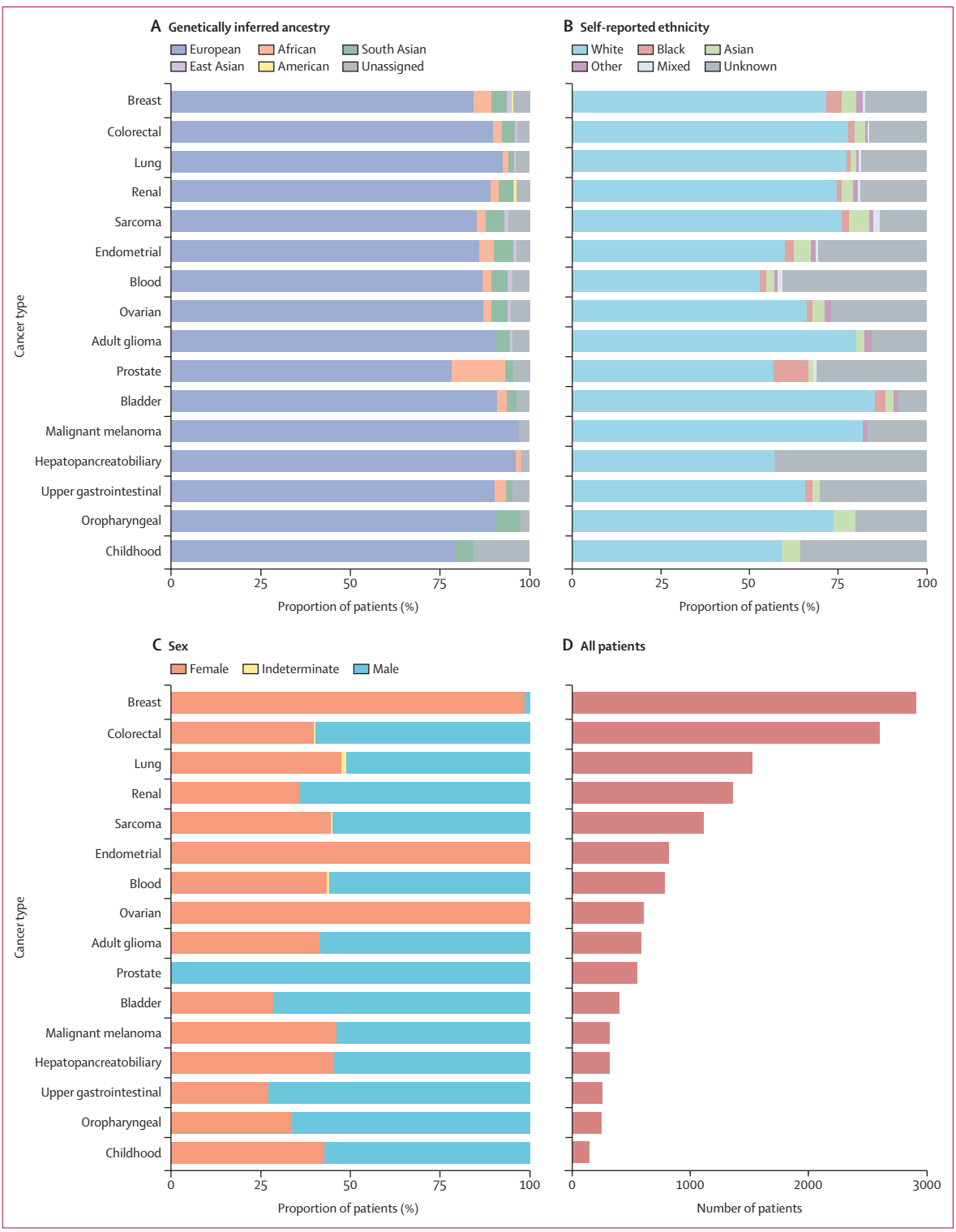

Figure 1: **Patient characteristics by cancer type in the 100 000 Genome Project**
(A) Percentage of patients assigned to each genetically inferred ancestral super-population group in the 100 000 Genomes Project Cancer Programme. (B) Percentage of patients by self-reported ethnicity. (C) Percentage of patients by sex. (D) Total number of patients. Cancers with fewer than 100 patients are not displayed. Sub-groupings with fewer than five patients for a cancer type are not included in percentage calculations. Tabular data are available in the appendix (pp 3–5).

mutations in any region of the genome that were smaller than 50 bp. We assessed whether TMB can be used as a proxy for tumour grade because there were high rates of missing data for the latter. We used ordinal logistic regression to predict grade as an ordered class outcome, TMB as a linear predictor, and cancer type as a covariate.

We conducted a sensitivity analysis on somatic mutations, restricted to targetable genes, that is those with potential treatment relevance. An additional sensitivity analysis was conducted in female patients with breast cancer to assess whether ancestry group effects on germline pathogenic variant detection were confounded by age. For this analysis, we created early (ie, age ≤46 years, n=394) and late (age ≥53 years, n=2020) onset groups and repeated the analysis described earlier.

We estimated population allele frequencies for identified candidate germline variants using data from the Genome Aggregation Database (gnomAD) version 3.1.[18] 100 000 Genomes Project ancestry groups were mapped to gnomAD super populations with the same names, except for the European ancestry group, which was mapped to Non-Finnish European in gnomAD.

We used a threshold for statistical significance of p<0·05 for all analyses, with adjustment for multiple testing where appropriate.

All statistical analyses were done using R (version 4.1.2), with regression analyses performed using the glmmTMB (version 1.2.2.3) R package (cran.r-project.org/web/packages/glmmTMB). Code for all analyses is available online.

Role of the funding source

The funder had no role in study design, collection, analysis, or interpretation of data, nor in the writing of the report or the decision to submit the paper for publication.

Results

14 775 patients with cancer were recruited between February, 2015, and December, 2018, into the 100 000 Genomes Project (figure 1). To assess ethnicity representation in the 100 000 Genomes Project, we compared the ethnic composition of 14 775 patients with cancer in the 100 000 Genomes Project (figure 2) with the ethnic composition of patients with cancer in England.[5] There was no evidence that non-White ethnic groups were under-represented when compared with national statistics (figure 2, appendix pp 6–7). The recruitment rate ratio for breast cancer was 2·2 (95% CI 1·6–3·0) for Black versus White women in the 100 000 Genomes Project compared with 0·8 (0·8–0·8) for Black versus White women in the national data (fold-change in rate ratios 2·7; 95% CI 2·0–3·7, p<0·0001), suggesting higher representation of Black women in the 100 000 Genomes Project than expected given the ethnicity-specific incidence rates in England. Compared with national rates, the 100 000 Genomes Project also had higher recruitment rates of Black versus White men with prostate cancer (fold-change in rate ratios 3·7; 1·8–7·5, p=0·0004), Black versus White men with bladder cancer (fold change in rate ratios 6·1; 95% CI 2·0–18·8, p=0.0016), and Asian versus White women with breast cancer (fold change in rate ratios 1·4; 1·2–1·7, p=0·0008).

We investigated ancestry differences in germline variant prioritisation. Patients with cancer types with fewer than five individuals, haematological cancers (n=788), childhood cancers (n=154), unknown primary carcinomas (n=84), patients with indeterminate sex (n=77), and patients missing somatic mutations in genes were excluded. Patients of admixed Latin American ancestry (n=35) were also excluded due to low sample numbers. Overall, 332 (2·4%) of 13 645 patients included in this analysis had one or more potentially pathogenic variant. 5608 (70·8%) of 7926 patients had at least one candidate germline variant (appendix p 8). The most common genes with potentially pathogenic variants were *BRCA2* in European and African ancestries and *BRCA1* in South Asian ancestries (further information on pathogenic variants in female patients with breast cancer are available in the appendix p 9).

Ancestry had a significant association with the likelihood of carrying a variant classified as potential pathogenic (likelihood ratio test p=0·0011). Potential pathogenic variants were identified in 23 (4·6%) of 500 patients with South Asian ancestry and 24 (5·3%) of 453 patients with African ancestry compared with 263 (2·2%) of 11 955 patients with European ancestry (appendix p 8). Compared with patients with European ancestries, significantly higher frequencies of potential pathogenic variants were found for patients with African (adjusted odds ratio [OR] 2·24; 95% CI 1·44–3·48, p=0·0003) and South Asian (adjusted OR 1·88; 1·21–2·93, p=0·0052) ancestries (figure 3A; appendix p 10). Furthermore, all the ancestry groups had significantly more candidate variants compared with the European ancestry group (p<0·0001; figure 3B; appendix p 11). Candidate variants were only called for versions 1.9 or later of the pipeline, so 5719 patients were not included in the analyses on candidate variants.

We assessed, using a sensitivity analysis, whether these differences might be linked to higher genetic diversity or heterozygous to homozygous ratio in some ancestry groups.[19] However, when adjusted for the total number of germline variants and for the heterozygous to homozygous ratio in addition to adjustment for age and cancer type, non-European ancestry groups consistently had more potential pathogenic variants and candidate variants than those in European ancestry groups and neither covariate was associated with the likelihood of having a potential pathogenic variant (appendix p 10).

In another sensitivity analysis, we investigated the effect of age on the likelihood of finding potential pathogenic variants in patients with breast cancer, the largest cancer type in the 100 000 Genomes Project. The most prevalent pathogenic variants were in the *BRCA1* and *BRCA2* genes,

For **analysis code** see https://gitlab.com/genomicsengland/Data_Diversity_Public/cancer-100kg-diversity-blog/-/tree/2024paper?ref_type=heads

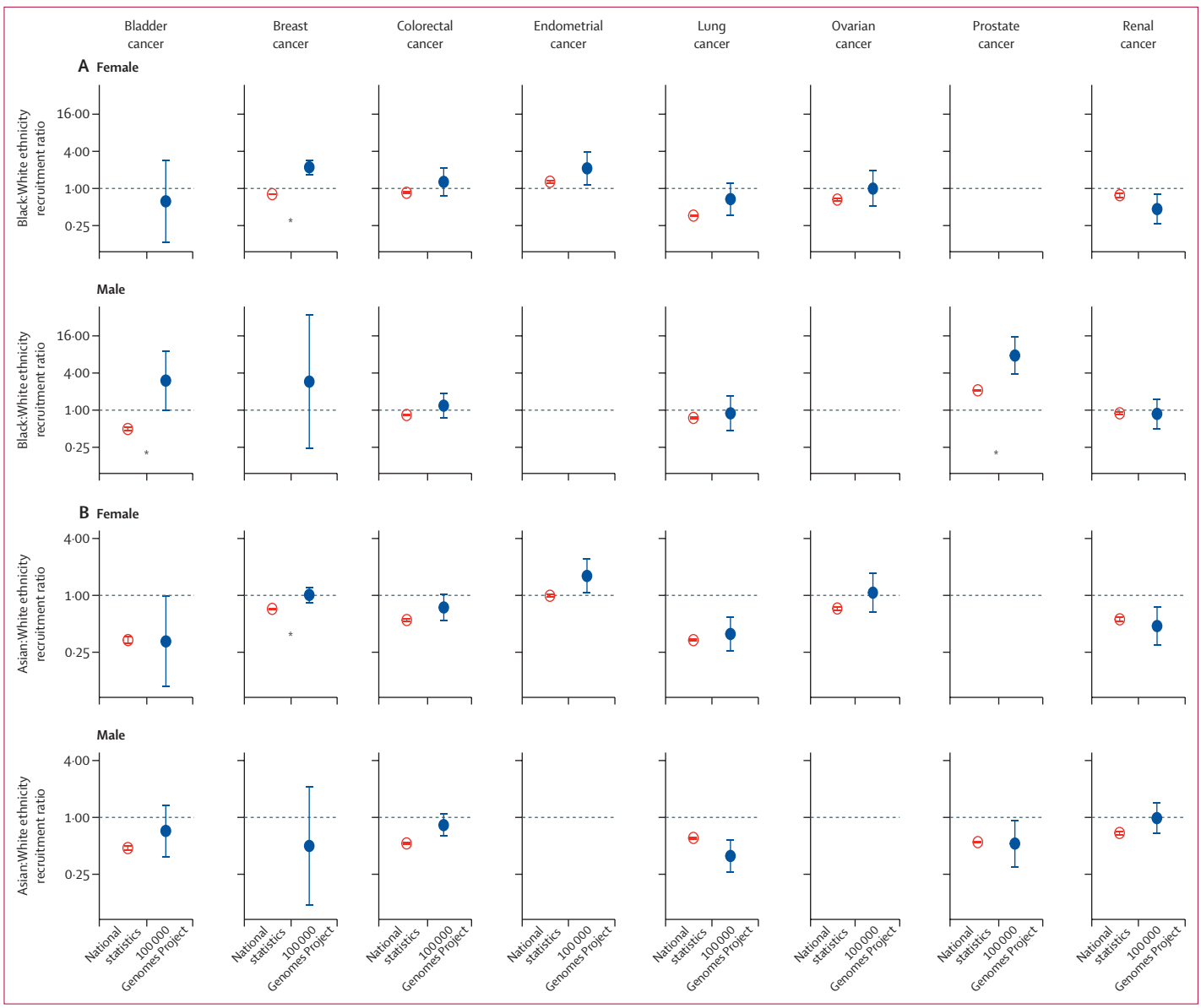

Figure 2: **Recruitment ratios for self-reported ethnicity (A) Black to White and (B) Asian to White**

Data are age-adjusted recruitment ratio for the 100 000 Genomes Project in blue and age-adjusted incidence ratios for England in red,[5] with 95% CIs. Comparisons for sub-categories with fewer than five patients in the 100 000 Genomes Project are omitted. *Statistically significant difference (p<0·05) in age-adjusted recruitment ratio after correction for multiple tests across cancer types.

which have previously been linked to earlier age at diagnosis.[20] The likelihood of identifying a pathogenic variant was associated with younger age across all cancers (p<0·0001). To separate the effect of age from ancestry, female patients with breast cancer were split into early-onset and late-onset groups. Only South Asian ancestry in the early-onset group and African ancestry in the late-onset group remained significantly associated with an increased likelihood of identifying a potential pathogenic variant in this small dataset (appendix p 12). For the East Asian ancestry group there were only 12 patients in the early-onset group and 17 patients in the late-onset group and none of them carried pathogenic variants.

We found that each non-European ancestry group had a proportion of candidate variants with allele frequencies greater than 2% (the common variant threshold used by the variant prioritisation pipeline; figure 4). 223 (25%) of 895 candidate variants in patients of African ancestry were common in the African gnomAD group, compared with 39 (0·5%) of 8362 in the European ancestry group, eight (1%) of 561 in the South Asian ancestry group, and three (2%) of 139 in the East Asian ancestry group (figure 4, appendix p 13).

For our analysis of ancestry differences in the prioritisation of somatic mutations, we compared somatic mutations in actionable genes for 13 645 patients

(table). We found that fewer somatic mutations in actionable genes were identified for patients of South Asian, East Asian, and African ancestry compared with patients of European ancestry when adjusting for sex and cancer type (likelihood ratio test p<0·0001). Compared with patients of European ancestry, patients of African ancestry had an incidence rate ratio of mutations in actionable genes of 0·89 (95% CI 0·80–0·99, p=0·028); patients of South Asian ancestry had an incidence rate ratio of 0·80 (0·73–0·89; p<0·0001), and patients of East Asian ancestry had an incidence rate ratio of 0·74 (0·60–0·92, p=0.0058; table; appendix p 22). Findings were consistent when restricting the analysis to targetable genes, except for African ancestry where the difference in the mutation count was not significant compared with European ancestry (appendix p 14). However, significantly fewer patients of African ancestry had any mutations in targetable genes, compared with those of European ancestry (p=0·025; appendix p 14).

We tested whether the differences in mutations in actionable genes are linked to ancestry differences in tumour characteristics. Although tumour stage was not associated with the number of somatic mutations in actionable genes (p=0·81), higher tumour grade was (table; p<0·0001). For some cancer types, there were ancestry-related differences in tumour grade (appendix pp 15, 23), with sarcoma tumour grade significantly lower in individuals of South Asian ancestry (p<0·0001), and breast cancer tumour grade higher in women of African ancestry (p=0·016), both compared with European ancestry. Ancestry still had a significant effect on the number of mutations in actionable genes after accounting for tumour grade (likelihood ratio test p=0·013, table [model 3]). However, grade was only available for around 60% of tumour samples (9278 samples across 8845 patients) resulting in a substantial loss of statistical power, and grade data were more likely to be missing for patients of African and unassigned ancestry and younger patients (appendix p 16).

Therefore, we also considered TMB which was available for the full sample and, when combined with cancer type, is predictive of grade (prediction accuracy 56·9%; appendix p 18). TMB was associated with an increased number of somatic mutations in actionable genes (p<0·0001; table). However, ancestry remained a significant predictor in the model adjusting for TMB (likelihood ratio test p=0·0006, table [model 2]). In addition, older age was associated with the number of somatic mutations in actionable genes. However, the effect of ancestry was still significant in models adjusted for age (data not shown, likelihood ratio test p=0·031).

Findings for somatic mutations across all genes (domains 1, 2, and 3 combined) were consistent with those for mutations in actionable genes (appendix pp 17, 22).

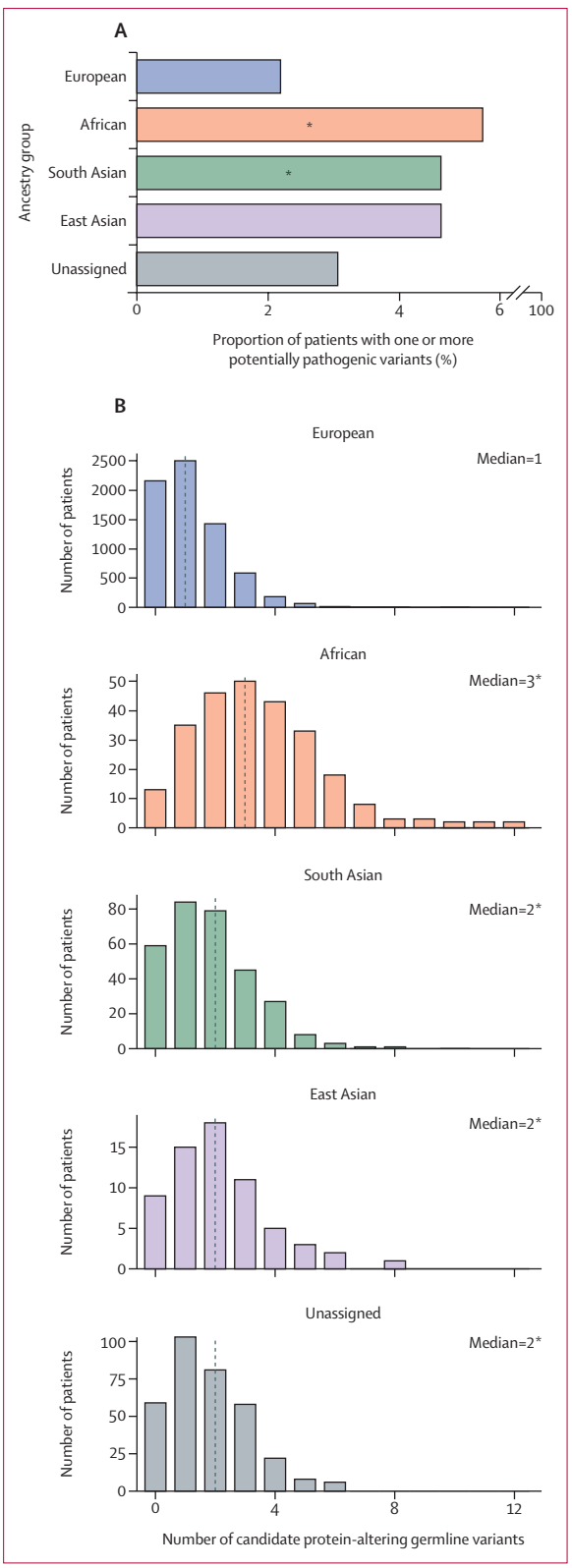

Figure 3: **(A) Percentage of potential pathogenic variants and (B) distribution of candidate variants by ancestry group**
The scales of the vertical axes in (B) vary by ancestry group. Full details are available in the appendix (pp 8, 10–11). *Statistically significant difference (p<0·05) between the number of variants for a given ancestry compared with individuals of European ancestry (model accounting for sex and cancer type).

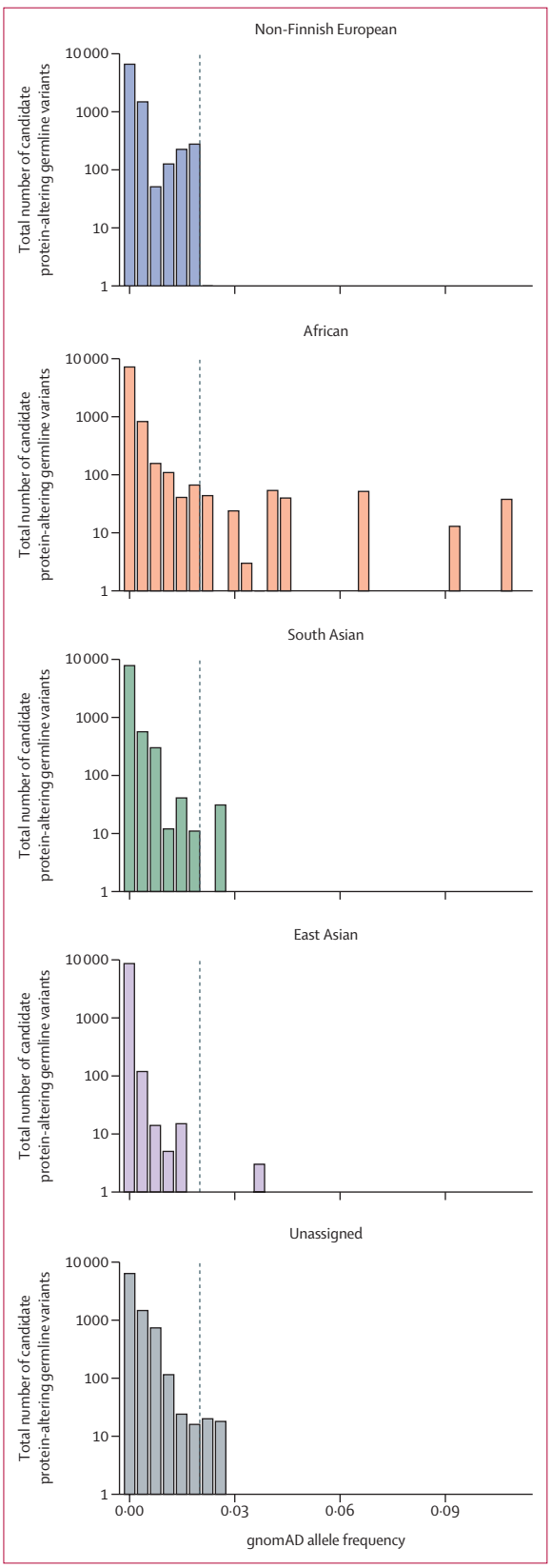

Figure 4: **Distribution of ancestry-specific allele frequencies in gnomAD for candidate germline variants, stratified by ancestry group**
The vertical line represents an allele frequency threshold of 2% to indicate common variants.

Discussion

We found that ethnicity in the 100 000 Genomes Project cancer programme is largely representative of the population of England, distinguishing it as a uniquely useful pan-cancer cohort for research on the role of ethnicity in diagnosis, prognosis, and care. We also identified statistically significant ancestry-related differences in the detection of both potentially pathogenic germline variants and somatic mutations in actionable genes.

To the best of our knowledge, this is the first study assessing ancestry differences in variant prioritisation for a national cancer cohort. Given its role as the pilot for nation-wide implementation of clinical whole-genome sequencing in cancer care, the representativeness of the 100 000 Genomes Project is of great importance. Other large studies assessing ancestry differences in cancer sequencing have suffered from under-representation of minority ethnic groups.[21] In the 100 000 Genomes Project, there was an over-recruitment of minority ethnicities relative to national incidence rates for some cancers.[22] As a limitation of our study, we cannot rule out the possibility that there are some recruitment biases affecting our findings, in particular for less common cancers. Our study does not allow conclusions to be drawn on the reasons for not finding under-representation of minority ethnic groups.

We found evidence that more germline variants in cancer susceptibility genes were classified as pathogenic variants for individuals of non-European ancestry. This finding was not explained by ancestry differences in genomic diversity or segregation mode. There is little previous evidence that cancer susceptibility variants are generally more frequent in non-European ancestry groups, except for specific founder effects.[23] We found that 25% of candidate germline variants in patients of African ancestry were common based on ancestry-specific reference data in gnomADv3.1. These variants are unlikely to be pathogenic. Some previous studies also found worse variant classification for individuals with non-European ancestry. For example, using data from 5026 patients from SEER in the USA, one study found variants of uncertain significance for 23·7% of White patients, but 44·5% of Black patients and 50·9% of Asian patients.[9] The difference in excess variants of uncertain significance versus potentially pathogenic variants is likely to be due to the different variant prioritisation and filtering approaches between studies. An excess of variants classified as potentially pathogenic variants seen in our data might have clinical disadvantages for patients with non-European ancestry in England, since they might make it more difficult to identify true pathogenic variants.

Future strategies to reduce the excess of variants classified as potentially pathogenic variants in these groups could be to improve annotations of variants in cancer susceptibility genes for diverse ancestries and to separate reference datasets into ancestry groups for

	Samples with mutations in actionable genes	Model 1 (N=13 645)		Model 2 (N=13 645)		Model 3 (N=8845)	
		Incidence rate ratio	p value	Incidence rate ratio	p value	Incidence rate ratio	p value
European	10 223/12 028 (85%)	1 (ref)	..	1 (ref)	..	1 (ref)	..
African	346/456 (76%)	0·89 (0·80–0·99)	0·028	0·91 (0·84–0·99)	0·025	0·94 (0·83–1·10)	0·39
South Asian	403/497 (81%)	0·80 (0·73–0·89)	<0·0001	0·92 (0·86–0·99)	0·030	0·89 (0·79–1·00)	0·053
East Asian	91/108 (84%)	0·74 (0·60–0·92)	0·0058	0·90 (0·77–1·00)	0·16	0·83 (0·66–1·10)	0·12
Unassigned	433/556 (78%)	0·83 (0·76–0·91)	0·0001	0·90 (0·84–0·97)	0·0044	0·89 (0·79–0·99)	0·036
Male sex	..	0·91 (0·87–0·95)	0·0001	1·0 (0·97–1·00)	0·83	0·85 (0·80–0·89)	<0·0001
Tumour mutation burden	..	..	..	1·00 (1·00–1·00)	<0·0001	..	..
Tumour grade	..	..	..	..	..	1·40 (1·33–1·43)	<0·0001

Data are mutations in actionable genes/number of pathogenic variants (%) or incidence rate ratio (95% CI) unless otherwise indicated. Model 1 was adjusted for sex and cancer type. Model 2 was adjusted for sex, cancer type, and tumour mutation burden. Model 3 was adjusted for sex, cancer type, and tumour grade. European ancestry and female sex were used as references. Model 3 was affected by missing data, with a 40% reduction in sample size compared with model 1 (see appendix p 17).

Table: **Association between genetically inferred ancestry group and the number of somatic mutations in actionable genes identified in tumour tissue**

filtering out common variants. Furthermore, it is important to continue to increase sample sizes and diversity of reference resources.

Conversely, fewer somatic mutations in potentially actionable genes were identified for individuals of non-European ancestry. This finding broadly aligns with findings from studies in the USA.[21] Using data from 45 000 patients with pan-cancer diagnoses from the MSK cohort, one study found targetable mutations in 30% of African ancestry patients versus 33% of European ancestry patients.[21] This study was different from ours in focusing on targetable somatic mutations and compared presence and absence rather than counts. Some genes, such as *EGFR*, show differential rates of somatic mutations in tumours across ancestry groups.[24,25] The under-representation of diverse ancestry groups in research studies could have favoured the development of targeted therapies for mutations in genes that are more frequent in European ancestries. To address this historic bias, future research should improve consideration of ancestry background across research stages in therapy development.

There are ethnicity-related differences for some tumour characteristics. For example, Black women in England present more frequently with triple negative breast cancer[26] and, in line with this finding, we observed that female patients with breast cancer of African ancestry had higher grade tumours. However, for patients with sarcoma, individuals of South Asian ancestry had significantly lower tumour grades. Other differences were not statistically significant, although this might be due to insufficient statistical power. When accounting for tumour grade or tumour mutational burden, the differences between European and other ancestry groups reduced, but the number of mutations in actionable genes remained consistently lower in African and South Asian ancestries. These findings suggest that the different detection rates of mutations in actionable genes might be partly mediated by differences in tumour characteristics.

However, as a limitation of our study, missing data made it difficult to assess the mediating role of tumour features more widely. For example, tumour grade was only available for around 60% of tumour samples.

Further study limitations include the small group sizes that resulted from that the intersection of ancestry and cancer types for some combinations. Therefore, most of the analyses adjusted for cancer types, but did not consider them separately as an outcome. The conclusions are unlikely to apply equally across all cancer types.

External validation is fundamentally challenging as we evaluated how a specific national whole-genome sequencing programme performs for patients with different ancestries. Other studies with whole-genome sequencing of patients with cancer not only had different designs, but also used different workflows for the identification of pathogenic variants or candidate variants and actionable tumour mutations. Moreover, such studies would need very large numbers of patients with cancer with whole-genome sequencing to enable the kinds of comparison that we undertook, and we are not aware of such a resource.

In conclusion, opportunities for further research will increase with the release of additional data from routine sequencing as part of the NHS Genomic Medicine Service. Our study underlines the need for this research, if we are to ensure equitable benefits from genomics in cancer prevention and care for diverse populations.

Contributors

MM, KK, STh, and MAB conceptualised the study. MAB and MM were involved in funding acquisition. AS and JA were involved in data collection. TN, STa, YC, JA, and STh carried out the data analysis. LM, MJS, KK, and MAB supervised the work. LM, KK, STa, MJS, TN, and AS interpreted the data. KK carried out the literature search. MM and MAB were involved in project administration. TN, KK, and MJS wrote the initial draft. All authors contributed to writing and editing and read and approved the manuscript. All authors had access to the data in the study.

TN, STa, and KK verified the data in the study. KK had final decision to submit the manuscript for publication.

Declaration of interests

Genomics England is a company wholly owned by the UK Department of Health and Social Care and was created in 2013 to introduce whole-genome sequencing into health care in conjunction with National Health Service England. All authors affiliated with Genomics England (TN, STa, YC, AS, JA, MM, MAB, LM, MJS, and KK) are, or were, salaried by Genomics England. TN became affiliated with e-Therapeutics after her contribution to this Article. MB received funding for a UCB PhD studentship and consulting fees from Altis Therapeutics and was on the Boards of Grey Wolf Therapeutics, Incyte, and Ipsen. MM received funding from The Health Foundation, consulting fees from Ellison Institute of Technology, support from Nature for travel, and has leadership roles with One HealthTech and Data Science for Health Equity. The other authors declare no competing interests.

Data sharing

The data supporting the findings of this study are available within the Genomics England Research Environment, a secure cloud workspace. Details on how to access data for this publication can be found at https://re-docs.genomicsengland.co.uk/pan_cancer_pub/. To access the genomic and clinical data within this research environment, researchers must first apply to become a member of either the Genomics England Research Network (previously known as the Genomics England Clinical Interpretation Partnership; www.genomicsengland.co.uk/research/academic) or a Discovery Forum industry partner (www.genomicsengland.co.uk/research/research-environment). The process for joining the Genomics England Research Network is described at www.genomicsengland.co.uk/research/academic/join-gecip. Deidentified participant data that have been made available to registered users include: alignments in BAM or CRAM format; annotated variant calls in VCF format; signature assignment; tumour mutational burden; sequencing quality metrics; summary of findings shared with the Genomic Lab Hubs; and secondary clinical data as described in this paper. The consent form can be found at https://www.genomicsengland.co.uk/assets/forms/Participant-consent-form-for-patients-with-cancer-or-suspected-cancer-C1.pdf.

Acknowledgments

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England (a wholly owned company of the UK Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the National Health Service (NHS) as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK, and the Medical Research Council have also funded research infrastructure. KK was supported by the EU under the Horizon 2020 Research and Innovation Programme (grant number 948561)

References

1 NHS England. NHS genomic medicine service. https://www.england.nhs.uk/genomics/nhs-genomic-med-service/ (accessed Feb 6, 2024).
2 Turnbull C, Scott RH, Thomas E, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* 2018; **361:** k1687.
3 Sosinsky A, Ambrose J, Cross W, et al. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nat Med* 2024; **30:** 279–89.
4 Alarcón Garavito GA, Moniz T, Déom N, Redin F, Pichini A, Vindrola-Padros C. The implementation of large-scale genomic screening or diagnostic programmes: a rapid evidence review. *Eur J Hum Genet* 2023; **31:** 282–95.
5 Delon C, Brown KF, Payne NWS, Kotrotsios Y, Vernon S, Shelton J. Differences in cancer incidence by broad ethnic group in England, 2013–2017. *Br J Cancer* 2022; **126:** 1765–73.
6 DeSantis CE, Miller KD, Goding Sauer A, Jemal A, Siegel RL. Cancer statistics for African Americans, 2019. *CA Cancer J Clin* 2019; **69:** 211–33.
7 Aizer AA, Wilhite TJ, Chen M-H, et al. Lack of reduction in racial disparities in cancer-specific mortality over a 20-year period. *Cancer* 2014; **120:** 1532–39.
8 Fry A, White B, Nagarwalla D, Shelton J, Jack RH. Relationship between ethnicity and stage at diagnosis in England: a national analysis of six cancer sites. *BMJ Open* 2023; **13:** e062079.
9 Kurian AW, Ward KC, Hamilton AS, et al. Uptake, results, and outcomes of germline multiple-gene sequencing after diagnosis of breast cancer. *JAMA Oncol* 2018; **4:** 1066–72.
10 Liu YL, Maio A, Kemel Y, et al. Disparities in cancer genetics care by race/ethnicity among pan-cancer patients with pathogenic germline variants. *Cancer* 2022; **128:** 3870–79.
11 Domchek SM, Yao S, Chen F, et al. Comparison of the prevalence of pathogenic variants in cancer susceptibility genes in Black women and Non-Hispanic White women with breast cancer in the United States. *JAMA Oncol* 2021; **7:** 1045–50.
12 Gov.uk. List of ethnic groups. https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups (accessed Nov 6, 2023).
13 Kousathanas A, Pairo-Castineira E, Rawlik K, et al. Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* 2022; **607:** 97–103.
14 Genomics England. Cancer analysis technical information document v1.11. May 17, 2019. https://files.genomicsengland.co.uk/forms/Cancer-Analysis-Technical-Information-Document-v1-11-main.pdf (accessed April 1, 2024).
15 Genomics England Research Environment User Guide. AggV2 functional annotation. https://re-docs.genomicsengland.co.uk/functional_annotation (accessed April 6, 2024).
16 Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet* 2019; **51:** 1560–65.
17 Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018; **46:** D1062–67.
18 Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020; **581:** 434–43.
19 Wall JD, Sathirapongsasuti JF, Gupta R, et al. South Asian medical cohorts reveal strong founder effects and high rates of homozygosity. *Nat Commun* 2023; **14:** 3377.
20 Peto J, Collins N, Barfoot R, et al. Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *J Natl Cancer Inst* 1999; **91:** 943–49.
21 Arora K, Tran TN, Kemel Y, et al. Genetic ancestry correlates with somatic differences in a real-world clinical cancer sequencing cohort. *Cancer Discov* 2022; **12:** 2552–65.
22 Gov.uk. Population of England and Wales. Dec 22, 2022. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/population-of-england-and-wales/latest (accessed Oct 27, 2023).
23 Neuhausen SL. Ethnic differences in cancer risk resulting from genetic variation. *Cancer* 1999; **86** (suppl)**:** 2575–82.
24 Shigematsu H, Lin L, Takahashi T, et al. Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *J Natl Cancer Inst* 2005; **97:** 339–46.
25 Carrot-Zhang J, Soca-Chafre G, Patterson N, et al. Genetic ancestry contributes to somatic mutations in lung cancers from admixed Latin American populations. *Cancer Discov* 2021; **11:** 591–98.
26 Gathani T, Reeves G, Broggio J, Barnes I. Ethnicity and the tumour characteristics of invasive breast cancer in over 116,500 women in England. *Br J Cancer* 2021; **125:** 611–17.