Genetic predisposition to cancer across people of different ancestries in Qatar: a population-based, cohort study

Mohamad Saad, Younes Mokrab*, Najeeb Halabi*, Jingxuan Shan, Rozaimi Razali, Khalid Kunji, Najeeb Syed, Ramzi Temanni, Murugan Subramanian, Michele Ceccarelli, Qatar Genome Programme Research Consortium†, Arash Rafii Tabrizi, Davide Bedognetti, Lotfi Chouchane*

Summary

Background Disparities in the genetic risk of cancer among various ancestry groups and populations remain poorly defined. This challenge is even more acute for Middle Eastern populations, where the paucity of genomic data could affect the clinical potential of cancer genetic risk profiling. We used data from the phase 1 cohort of the Qatar Genome Programme to investigate genetic variation in cancer-susceptibility genes in the Qatari population.

Methods The Qatar Genome Programme generated high-coverage genome sequencing on DNA samples collected from 6142 native Qataris, stratified into six distinct ancestry groups: general Arab, Persian, Arabian Peninsula, Admixture Arab, African, and South Asian. In this population-based, cohort study, we evaluated the performance of polygenic risk scores for the most common cancers in Qatar (breast, prostate, and colorectal cancers). Polygenic risk scores were trained in The Cancer Genome Atlas (TCGA) dataset, and their distributions were subsequently applied to the six different genetic ancestry groups of the Qatari population. Rare deleterious variants within 1218 cancer susceptibility genes were analysed, and their clinical pathogenicity was assessed by ClinVar and the CharGer computational tools.

Findings The cohort included in this study was recruited by the Qatar Biobank between Dec 11, 2012, and June 9, 2016. The initial dataset comprised 6218 cohort participants, and whole genome sequencing quality control filtering led to a final dataset of 6142 samples. Polygenic risk score analyses of the most common cancers in Qatar showed significant differences between the six ancestry groups ($p<0·0001$). Qataris with Arabian Peninsula ancestry showed the lowest polygenic risk score mean for colorectal cancer (–0·41), and those of African ancestry showed the highest average for prostate cancer (0·85). Cancer-gene rare variant analysis identified 76 Qataris (1·2% of 6142 individuals in the Qatar Genome Programme cohort) carrying ClinVar pathogenic or likely pathogenic variants in clinically actionable cancer genes. Variant analysis using CharGer identified 195 individuals carriers (3·17% of the cohort). Breast cancer pathogenic variants were over-represented in Qataris of Persian origin (22 [56·4%] of 39 *BRCA1/BRCA2* variant carriers) and completely absent in those of Arabian Peninsula origin.

Interpretation We observed a high degree of heterogeneity for cancer predisposition genes and polygenic risk scores across ancestries in this population from Qatar. Stratification systems could be considered for the implementation of national cancer preventive medicine programmes.

Funding Qatar Foundation.

Introduction

The risk of developing cancer varies according to race, ethnicity, or ancestry.[1] Countries in the Middle East have been experiencing an alarming increase in cancer rates in the past decade.[2] In Qatar, cancer is the nation's second most prevalent non-communicable disease, and the prevalence is projected to increase because of a combination of ageing and population growth.[3] Numerous disease-associated gene variants, including those related to cancer, show substantial diversity in ancestral and derived allele frequencies among different populations. However, disparities in the genetic risk of cancer between ancestry groups remain poorly defined. To our knowledge, Arabian populations, despite their diversity, have not been included in international genome or cancer consortia. The population structure of Arabs might result in the emergence of founder variants that could influence the development or progression of cancer.[4]

Next-generation DNA sequencing is increasingly being considered as a core component of precision medicine because of its rapid developments and because of the potential of whole genome and exome sequencing in predicting genetic predisposition to many diseases.[5] As decreasing costs make next-generation sequencing increasingly affordable, the search for germline variations in cancer susceptibility genes will move from single-gene approaches to genome-wide analyses. Consequently, the targeted population will expand from at-risk individuals of families with cancer to individuals from the general population.

Lancet Oncol 2022; 23: 341–52

Published Online February 9, 2022 https://doi.org/10.1016/S1470-2045(21)00752-X

See Comment page 318

*Contributed equally as first authors

†Members of the Qatar Genome Programme Research Consortium are listed in appendix 1 (pp 21–23)

Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar (M Saad PhD, K Kunji MSc); **Department of Human Genetics** (Y Mokrab PhD, R Razali PhD), **Applied Bioinformatics Core, Integrated Genomics Services, Research Branch** (N Syed MSc), **and Cancer Research Department, Research Branch** (D Bedognetti MD), **Sidra Medicine, Doha, Qatar; Genetic Intelligence Laboratory** (N Halabi PhD, J Shan PhD, M Subramanian PhD, Prof A Rafii Tabrizi MD, Prof L Chouchane DSc), **Weill Cornell Medicine-Qatar, Doha, Qatar; Department of Genetic Medicine** (J Shan, Prof A Rafii Tabrizi, Prof L Chouchane), **and Department of Microbiology and Immunology** (M Subramanian, Prof L Chouchane), **Weill Cornell Medicine, New York, NY, USA; Janssen Research and Development, Paris, France** (R Temanni PhD); **Department of Electrical Engineering and Information Technology (DIETI), University of Naples "Federico II", Naples, Italy** (M Ceccarelli PhD); **Biogem, Istituto di Biologia e Genetica Molecolare, Ariano Irpino, Italy** (M Ceccarelli)

Correspondence to: Professor Lotfi Chouchane, Weill Cornell Medicine-Qatar, Education City, Qatar Foundation, Doha, Qatar **loc2008@med.cornell.edu**

See Online for appendix 1

Research in context

Evidence before this study

We searched PubMed for published articles focusing on cancer-gene germline variations in Middle Eastern populations, using various combinations of the search terms "germline cancer gene variation" OR "polygenic risk scores" AND "Arab population" OR "Middle Eastern population". We also screened preprint servers such as medRxiv for related articles. We did not find any reports describing germline variations in cancer-related genes in Middle Eastern populations.

Added value of this study

This study is, to the best of our knowledge, the first large genomic analysis of cancer-gene germline variations in a Middle Eastern population, which constitutes a valuable resource to capture cancer gene variation in ancestry-based Middle Eastern populations. We describe the development of the first polygenic risk scores for three common cancers (breast, prostate, and colorectal cancers) in the Qatari population. Additionally, we showed that these polygenic risk scores are ancestry specific; even within populations confined to a single geographical location, such as Qatar, polygenic risk scores can vary significantly. Furthermore, our analysis of rare cancer-gene variants revealed several characteristics unique to the native Qatari population. According to our data, none of the Qataris with Arabian Peninsula ancestry was found to be carriers of breast or ovarian cancer variants. By contrast, these variants were over-represented among Qataris of Persian ancestry.

Implications of all the available evidence

A high degree of heterogeneity of cancer predisposition genes and polygenic risk scores across ancestries has been observed in the Qatari population. Population stratification should be considered in the implementation of national cancer prevention programmes. In comparison with other human populations, the Qatari population possessed a high prevalence of certain rare deleterious mutations in clinically actionable cancer genes. The high consanguinity rate of the Qatari population might be an important contributing factor in the over-presence of these uncommon variants. Our data constitute a call to action to complement the use of the ClinVar database with other computational predictors (eg, CharGer) to enhance the actionability of rare cancer-gene pathogenic or likely pathogenic variants.

Aside from certain rare variants, which cause inherited autosomal conditions such as *BRCA*-related hereditary breast cancer, ovarian cancer, and Lynch syndrome, for which early identification and intervention facilitate clinical actionability and have a positive impact on public health,[6] a large proportion of the genetic risk of cancer stems from the interactive effect of multiple low-risk and modest-risk variants. Combining the weighted values of these variants results in the generation of polygenic risk scores.[7] Studies have started to report the clinical utilities of polygenic risk scores,[8,9] resulting in their increased commercial availability.[10] The portability of polygenic risk scores to diverse populations needs to be assessed to avoid further health disparities.

For more on the **Qatar Biobank** see http://www.qatarbiobank.org.qa

The Qatar Genome Programme is a population-based project launched by Qatar Foundation to generate a large whole genome sequencing dataset. In combination with comprehensive clinical and biological information collected by the Qatar Biobank,[11] the Qatar Genome Programme is aiming to investigate the genetic variation in disease-causing genes, which will pave the way for precision medicine in Qatar.[11,12] In this study, using the whole genomes of 6142 individuals from the Qatar Genome Programme phase 1 cohort, we aimed to estimate the polygenic cancer risk in the Qatari population through polygenic risk score analyses for the most common cancers in Qatar. We extended our work by defining the landscape of rare deleterious variants of 1218 cancer-susceptibility genes and assessing their predicted clinical pathogenicity.

Methods

Study participants and dataset

The initial dataset comprised 6218 Qatar Biobank participants, recruited as part of an ongoing, longitudinal, population-based study aiming to recruit 60 000 individuals from the Qatari population. Detailed information about the cohort, phenotypic data, and collection of samples by the Qatar Biobank have been described elsewhere.[11] Briefly, participants were recruited either by personal recommendations of family and friends or via social media and the Qatar Biobank website. Participants attended an assessment session at the Qatar Biobank facilities, Doha, in which physical measurements were taken. Standardised paper questionnaires reporting information about lifestyle, diet, and medical history were filled by each participant. Biological samples (blood, saliva, and urine) were provided and stored at –80°C in liquid nitrogen.

Written, informed consent was obtained from all Qatar Biobank participants before their participation, and the study was approved by the Hamad Medical Corporation Ethics Committee and the Qatar Biobank institutional review board.

Whole genome sequencing

DNA was extracted from peripheral blood with the automated QIASymphony SP instrument according to the Qiagen MIDI kit protocol's recommendations (Qiagen; Hilden, Germany). Whole genome libraries were prepared from 150 ng of DNA with the Illumina TruSeq DNA Nano

kit (San Diego, CA, USA). Genomic libraries were sequenced on HiSeq X Ten (Illumina, San Diego, CA, USA) following the manufacturer's recommended protocol to achieve a minimum average coverage of 30×. Library construction and sequencing was done at the Sidra Clinical Genomics Laboratory, Doha, Qatar. Quality control of Fastq files was done with FastQC (version 0.11.2). Reads were then aligned to the GRCh37 (hs37d53) reference genome by use of bwa.kit (version 0.7.12). Quality control on mapped reads was done with the Picard toolkit ([CollectWgsMetrics]; version 1.117). Variant calling was done jointly on all samples with Genome Analysis Toolkit (GATK) 3.4 best practices. Details of bioinformatics analyses and data quality control are provided in appendix 1 (p 11, 18–20).

The ancestry groups of the Qatari population were established with the same cohort as previously described.[13] Extensive analysis was done with Admixture, Principal Component Analysis, Fst, and F3 statistics, in which data from more than 3900 individuals were incorporated from various public reference datasets, including the 1000 Genomes Project, Human Origin Project, Greater Middle Eastern study, and previously published Qatari genomes. These analyses were done on autosomal single-nucleotide variants. Mitochondrial and Y chromosome haplogroups were also analysed to confirm the nature of the ancestry groups. One outlier was removed because its genetic signature matched the European ancestry of the 1000 Genomes. Detailed information about the Qatari population structure has been described elsewhere.[13,14] To assess differences in allele frequencies of cancer-gene variants across ancestry groups of the Qatar Genome Programme, the heterozygosity rate using the common variants (minor allele frequency >0·01) within cancer-susceptibility genes was computed with PLINK (version 1.09).

Statistical analysis

For the analysis of common variants associated with cancers, we focused on the three most common cancers in Qatar (breast, prostate, and colorectal). The discriminative power of nine existing polygenic risk scores was evaluated: five for breast cancer, three for prostate cancer, and one for colorectal cancer. Polygenic risk scores were downloaded from the Polygenic Risk Score Catalog. Polygenic risk scores were computed with the PLINK (version 1.09) '--score' command. The predictive performance of the nine polygenic risk scores was assessed in The Cancer Genome Atlas (TCGA) imputed dataset[15] and then the best polygenic risk score was applied to the Qatar Genome Programme cohort, stratified by ancestry. The best polygenic risk scores were compared between women and men, and between young and old individuals (age <40 years *vs* ≥40 years) using *t* test. Because the TCGA dataset does not contain cancer-free controls, evaluation of polygenic risk score performance for a cancer type was compared between the cancer type and for all remaining cancers combined (pseudo controls). Details about polygenic risk score selection and computing are provided in appendix 1 (pp 4, 8, 12).

We then did an analysis of rare pathogenic or likely pathogenic variants within cancer susceptibility genes. This analysis focused on a list of cancer-susceptibility genes to identify rare pathogenic or likely pathogenic variants that are differentially prevalent between our cohort and various publicly available datasets. The list of cancer-susceptibility genes was constructed by combining different sources (appendix 1 pp 12–13). We defined three different gene classes (classes 1, 2, and 3). Class 1 comprises genes with specific clinical actionability, which refers to a specific screening programme, family genetic counselling, or medical or surgical prevention. Class 2 comprises genes that are likely to be actionable. Class 3 comprises genes that are not yet actionable, but are associated with cancer. For the selection of rare genetic variants, we carried out the following steps. Starting from the quality processed VCF file containing 6142 samples, we obtained all variants that overlapped with our cancer gene list on the basis of gene coordinates from GENCODE 33,[16] and annotated variants using the Variant Effect Predictor.[17] The annotation included various functional information from the Human Genome Mutation Database (version 2018.2),[18] ClinVar (downloaded version on March 2, 2020),[19] and CharGer.[20] CharGer is a software that provides implementation of the American College of Medical Genetics and Genomics rules for germline pathogenic classification, and has been used by TCGA to annotate germline variants.[21] We included in our final variant list any variant annotated as pathogenic or likely pathogenic in either ClinVar or CharGer without any conflicting interpretations at a maximum allele frequency less than 1%. Concordance rates between ClinVar and CharGer in the identification of pathogenic or likely pathogenic variants were calculated. The variant burden in specific gene groups was compared between different subpopulations with Fisher's exact test, with a false-discovery rate controlled with the Benjamini–Hochberg procedure. A corrected p value less than 0·05 was considered statistically significant. Relatedness between individuals carrying rare variants was calculated with the Kstat tool. The proportion of unrelated individuals was computed as the percentage of kinship coefficients less than 0·0325. Additional details of the list of cancer-susceptibility genes used, their clinical actionability stratification and computing for the selection of rare genetic variants are provided in appendix 1 (pp 12–15). Variants are referred to with dbSNP reference rs numbers, and as "chromosome:position:reference allele:alternate allele" if rs numbers are not available.

Role of the funding source

The sponsor of study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

For more on **FastQC** see https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

For the **bwa.kit scripts** see https://github.com/lh3/bwa/tree/master/bwakit

For the **Picard toolkit** see https://gatk.broadinstitute.org/hc/en-us

For the **Genome Analysis Toolkit 3.4 best practices** see https://software.broadinstitute.org/gatk/documentation/article?id=3238

For the **Polygenic Risk Score Catalog** see https://www.pgscatalog.org

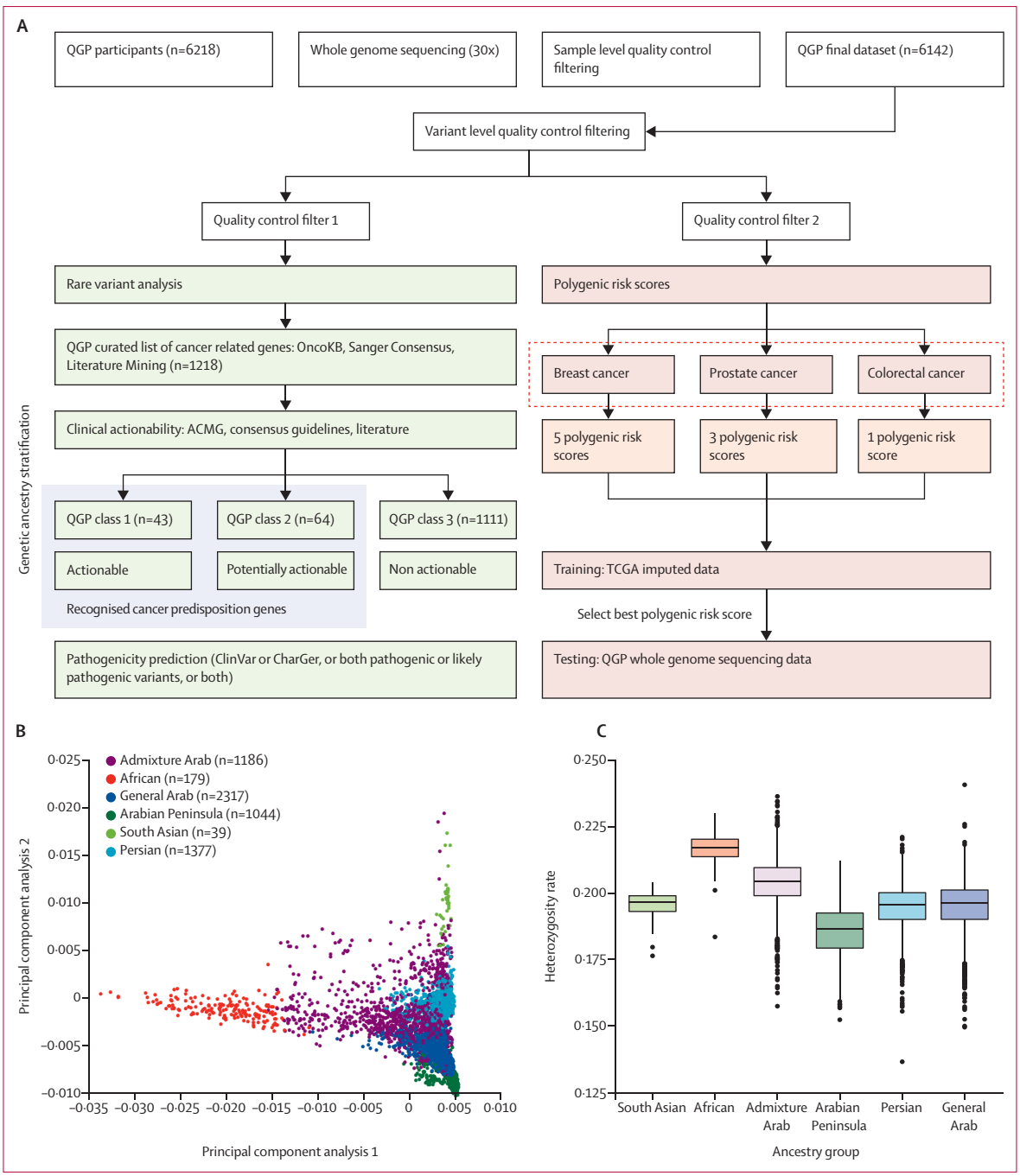

Figure 1: **Flowchart of the study and the population genetic structure of the Qatari people**
(A) Flowchart of the study design showing polygenic risk score and cancer rare variant analysis. (B) Population structure analysis showing the six major ancestry groups inferred by principal component analysis. (C) Heterozygosity rates per individual, stratified by ancestry groups, calculated with common variants (minor allele frequency >0·01) in cancer-related genes. ACMG=American College of Medical Genetics and Genomics. QGP=Qatar Genome Programme. TCGA=The Cancer Genome Atlas.

Results

The Qatar Genome Programme cohort comprised 6218 participants, recruited between Dec 11, 2012, and June 9, 2016, and the study was done from Oct 19, 2017, to July 7, 2021. Whole genome sequencing quality control filtering led to a final dataset of 6142 samples (figure 1A), stratified into six distinct ancestry groups: individuals with a general Arab origin (n=2317; 37·7% of the Qatar Genome Programme dataset), a more eastern or Persian origin (n=1377; 22·4%), an admixture of the different Arab subpopulations (n=1186; 19·3%), an Arabian Peninsula origin (n=1044; 17·0%), African origin (n=179;

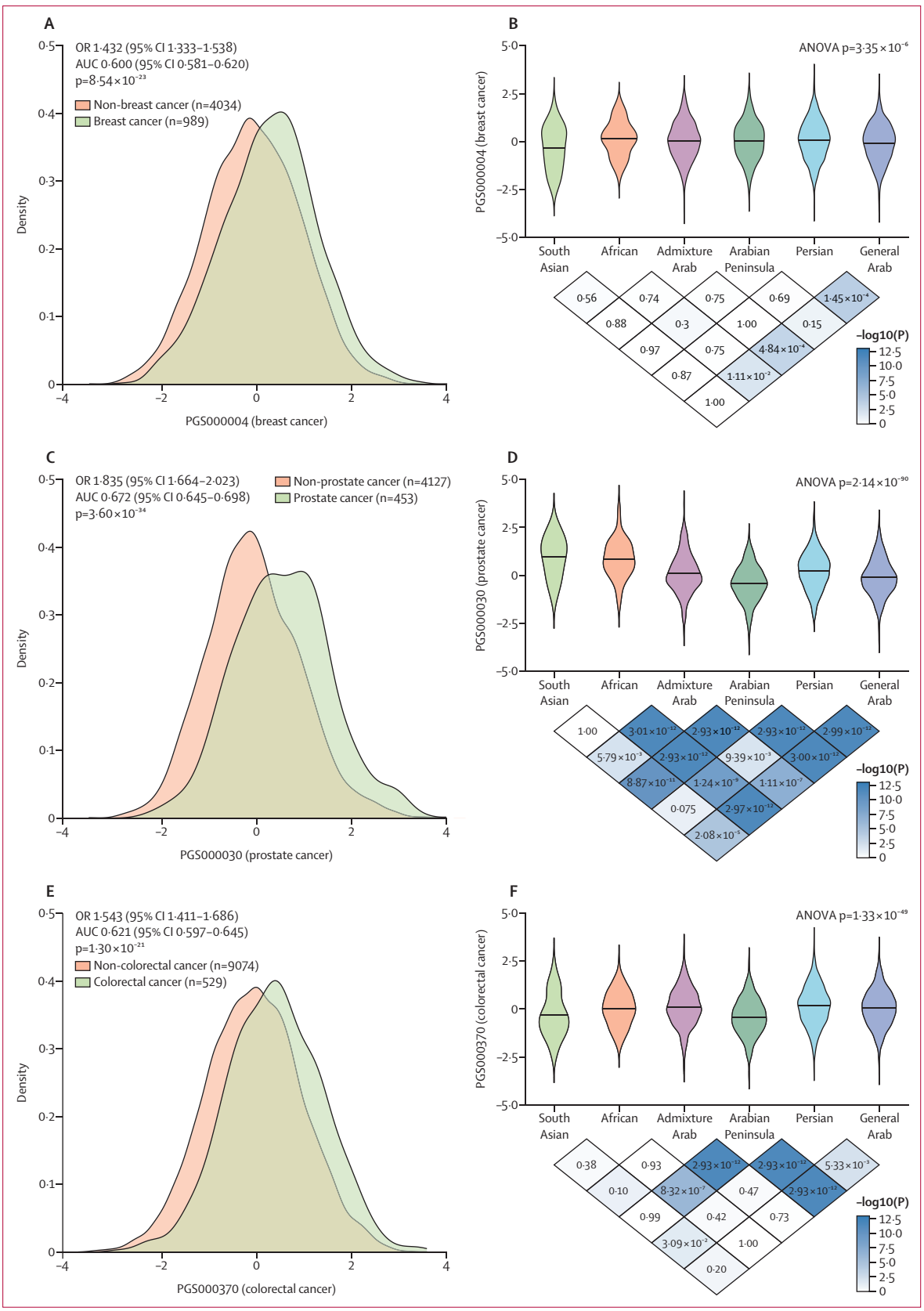

Figure 2: **Polygenic risk score distribution in TCGA and QGP datasets for the best polygenic risk scores of the three most common cancers in Qatar**

(A) Density plot of the best polygenic risk score for individuals with breast cancer (PGS000004) versus those with non-breast cancer in The Cancer Genome Atlas (TCGA). (B) Distribution of polygenic risk score for individuals with breast cancer in the Qatar Genome Programme (QGP) ancestry groups. (C) Density plot of the best polygenic risk score for individuals with prostate cancer (PGS000030) versus those with non-prostate cancer in TCGA. (D) Distribution of polygenic risk score for individuals with prostate cancer in QGP ancestry groups. (E) Density plot of the best polygenic risk score for individuals with colorectal cancer (PGS000370) versus those with non-colorectal cancer in TCGA. (F) Distribution of polygenic risk score for individuals with colorectal cancer in QGP ancestry groups. Tukey's post-hoc p values for differences between pairs of groups are shown in panels B, D, and F. Black horizontal lines on the violin plots (B, D, and F) are the polygenic risk score means. OR=odds ratio per 1 SD increase. AUC=area under the receiver operating curve. ANOVA was performed for comparison between the six QGP ancestry groups.

2·9%), and a small number of individuals with a South Asian origin (n=39; 0·6%; figure 1B).

The group of Arabian Peninsula Qataris showed the lowest heterozygosity rate, whereas those with African ancestry showed the highest heterozygosity rate (figure 1C). Statistically significant differences in the heterozygosity rate between the ancestry groups were observed (p<0·0001).

The nine polygenic risk scores for breast, prostate, and colorectal cancers, trained in the TCGA imputed dataset, are shown in appendix 1 (p 9). For breast cancer, PGS000004 showed the best performance (odds ratio for a 1 SD increase [OR_{1SD}] 1·432 [95% CI 1·333–1·538]; area under the receiver operating curve [AUC] 0·600, 95% CI 0·581–0·620; p=$8·54\times10^{-23}$; figure 2A; appendix 1 p 2). For prostate cancer, the best performance was seen with PGS000030 (OR_{1SD} 1·835 [95% CI 1·664–2·023]; AUC 0·672, 0·645–0·698; p=$3·59\times10^{-34}$; figure 2C; appendix 1 p 3). The colorectal cancer polygenic risk score (ie, PGS000370) performed relatively well (OR_{1SD} 1·543 [95% CI 1·411–1·686], AUC 0·621, 0·597–0·645; p=$1·30\times10^{-21}$; figure 2E). ANOVA testing showed significant differences between the six ancestry groups for the three cancers (p<0·0001; figure 2B, D, F). Differences were more pronounced, in terms of statistical significance, for colorectal and prostate cancers. For the following comparison, we omitted the South Asian ancestry group because of its small sample size. For breast cancer, the general Arab ancestry group had the lowest polygenic risk score average (–0·097). Significant differences were observed between individuals of general Arab origin versus those of Persian origin, those of general Arab origin versus an admixture Arab subpopulations, and those of general Arab origin versus those with an African ancestry (figure 2B). For prostate cancer, the Arabian Peninsula ancestry group showed the lowest polygenic risk score average (–0·43) and individuals with an African ancestry showed the highest average (0·85). All groups differed significantly from each other (figure 2D). For colorectal cancer, individuals of Arabian Peninsula origin showed the lowest polygenic risk score mean (–0·41), followed by those of African origin (0·018). The Arabian Peninsula ancestry group differed significantly from all remaining groups (figure 2F). PGS000004, PGS000030, and PGS000370 were compared between young and old individuals (age <40 years *vs* ≥40 years), but the risk scores did not differ significantly (p=0·48 for PGS000004, p=0·091 for PGS000030, and p=0·055 for PGS000370). Moreover, the colorectal cancer polygenic risk score, PGS000370, did not differ significantly between women and men (p=0·49). Breast and prostate cancer polygenic risk scores were not compared between sexes because they were evaluated in only one sex.

Rare deleterious variants of 1218 cancer-susceptibility genes were analysed. The distribution of the pathogenic and likely pathogenic variants of the three cancer-gene classes in the six ancestry groups of the Qatar Genome Programme are shown in the table and figure 3. Pathogenic and likely pathogenic variants were present in 23 (53·5%) of 43 class 1 genes, in 29 (45·3%) of 64 class 2 genes, and in 113 (10·2%) of 1111 class 3 genes (table; figure 3A). Two individuals carried more than one pathogenic or likely pathogenic variant in class 1 genes, and five individuals carried more than one pathogenic or likely pathogenic variant in class 2 genes. These seven individuals carried two variants out of 11 unique variants, which means that three variants were shared by more than one individual (appendix 2). The distribution of the pathogenic or likely pathogenic carrier frequencies among the six ancestry groups of the Qatar population showed an ancestry-dependent pattern (figure 3B).

We identified 333 (20·7%) deleterious pathogenic or likely pathogenic variants by either ClinVar or CharGer in 1273 unique individuals (table). Based on ClinVar, 76 individuals (comprising 1·2% of the cohort) were found to be carriers of a pathogenic or likely pathogenic variant in clinically actionable genes (22 pathogenic or likely pathogenic variants in 11 class 1 genes). Additionally, 23 pathogenic or likely pathogenic variants in 15 potentially actionable genes (class 2) were found in 61 individuals (1·0% of the cohort). Interestingly, CharGer predictions revealed additional pathogenic or likely pathogenic variants (37 variants in 19 class 1 genes and 39 variants in 22 class 2 genes), leading to an increase in the total number of individuals carrying these variants: 195 individuals (3·2% of the cohort) carrying a pathogenic or likely pathogenic variant in class 1 genes and 175 (2·9%) carrying a pathogenic or likely pathogenic variant in class 2 genes. Consequently, of the 6142 Qataris in the current cohort of the Qatar Genome Programme, 370 (6·0%) were found to be carriers of a pathogenic or likely pathogenic variant in clinically actionable or potentially actionable cancer genes (class 1 and class 2). For class 3 genes, ClinVar analysis showed the presence of 121 pathogenic or likely pathogenic variants in 76 genes carried by 796 (13·0%) individuals, whereas CharGer identified an additional 91 pathogenic or likely pathogenic variants in 47 genes carried by an additional 192 (3·1%) individuals. The assessment of the concordance rate of detection of pathogenic or likely pathogenic variants by ClinVar and CharGer is shown in appendix 1 (pp 10, 16–17).

The search for pathogenic or likely pathogenic variants in the most clinically actionable genes (class 1) showed that many variants (54 [91·5%] of 59) were more frequent in the Qatar Genome Programme dataset than in public datasets (figure 4A; appendix 2), including the following genes: *EPCAM* (two variants in 29 individuals), *MUTYH* (five variants in 24 individuals), *BRCA1* (eight variants in 21 individuals), and *BRCA2* (eight variants in 18 individuals; figure 4B).

Variants within breast cancer genes (*BRCA1, BRCA2, ATM*, *CHEK2*, and *PALB2*) and Lynch syndrome genes

See Online for appendix 2

	Individual carriers*	Number of variants	Number of genes	Individuals with more than one variant	Admixture Arab	African	General Arab	Arabian Peninsula	South Asian	Persian
Class 1										
Pathogenicity prediction										
All pathogenic and likely pathogenic	195 (3·2%)	59	23	2	41	6	80	23	1	44
ClinVar										
Pathogenic or likely pathogenic	76 (1·2%)	22	11	1	15	0	33	20	0	8
Pathogenic	73 (1·2%)	20	9	1	13	0	32	20	0	8
Likely pathogenic	3 (0·1%)	2	2	0	2	0	1	0	0	0
CharGer only										
Pathogenic or likely pathogenic	120 (2·0%)	37	19	0	26	6	48	3	1	36
Pathogenic	30 (0·5%)	16	10	0	5	2	7	1	0	15
Likely pathogenic	90 (1·5%)	21	13	0	21	4	41	2	1	21
Class 2										
Pathogenicity prediction										
All pathogenic and likely pathogenic	175 (2·9%)	62	29	5	45	10	70	10	1	39
ClinVar										
Pathogenic or likely pathogenic	61 (1·0%)	23	15	0	18	1	23	8	1	10
Pathogenic	58 (0·9%)	21	14	0	15	1	23	8	1	10
Likely pathogenic	3 (0·1%)	2	2	0	3	0	0	0	0	0
CharGer only										
Pathogenic or likely pathogenic	114 (1·9%)	39	22	5	27	9	47	2	0	29
Pathogenic	21 (0·3%)	10	6	2	4	0	16	0	0	1
Likely pathogenic	93 (1·5%)	29	17	3	23	9	31	2	0	28
Class 3										
Pathogenicity prediction										
All pathogenic and likely pathogenic	972 (15·8%)	212	113	101	192	18	399	198	11	154
ClinVar										
Pathogenic or likely pathogenic	796 (13·0%)	121	76	50	139	11	345	186	5	110
Pathogenic	421 (6·9%)	74	51	13	81	3	177	104	3	53
Likely pathogenic	403 (6·6%)	47	36	12	65	8	180	89	2	59
CharGer only										
Pathogenic or likely pathogenic	192 (3·1%)	91	47	37	55	7	61	14	6	49
Pathogenic	39 (0·6%)	27	9	12	8	3	16	2	1	9
Likely pathogenic	154 (2·5%)	65	42	25	47	4	45	12	5	41

Data are n, unless otherwise indicated. *Percentage corresponds to proportion of individual carriers out of all individuals in the Qatar Genome Programme.

Table: **Distribution of the pathogenic and likely pathogenic variants of the three cancer-gene classes in the six ancestry groups of the Qatar Genome Programme dataset with ClinVar and CharGer**

(*MLH1, MSH2, MSH6, PMS2,* and *EPCAM*) showed an ancestry-dependent pattern in the Qatari population. Highly significant differences in carrier frequencies of breast cancer and Lynch syndrome gene pathogenic or likely pathogenic variants were seen between Qataris of Arabian Peninsula origin and those of Persian origin (appendix 1 p 5). Of the 39 *BRCA1/BRCA2* variant carriers, 22 (56·4%) individuals were of Persian origin (figure 4C). Interestingly, none of the Qataris of Arabian Peninsula origin was found to be carriers of any pathogenic or likely pathogenic variants on *BRCA1/BRCA2* or other breast cancer-associated genes, including *ATM*, *CHEK2*, and *PALB2*. In contrast to breast cancer, Lynch syndrome variant carriers were over-represented in the Arabian Peninsula origin group (20 [59·0%] individuals out of the 34 carriers) and absent in the Persian origin group (appendix 1 p 5). A significant difference in the class 2 pathogenic or likely pathogenic carrier frequency was also observed for individuals with Arabian Peninsula origin (the ratio of allele frequency in Arabian Peninsula individuals to the allele frequency in the whole population was 0·321 [0·009/0·028], p=0·0001; appendix 1 p 5). All 333 variant frequencies were compared between women and men (appendix 2). The burden of breast cancer genes and Lynch syndrome genes was also compared between women and men. For breast cancer genes, the burden of rare pathogenic variants was 0·6% (22 out of 3460) in women and around 1·1% (30 out of 2682) in men. Men carried 1·76-times more breast cancer gene variants than women. For Lynch

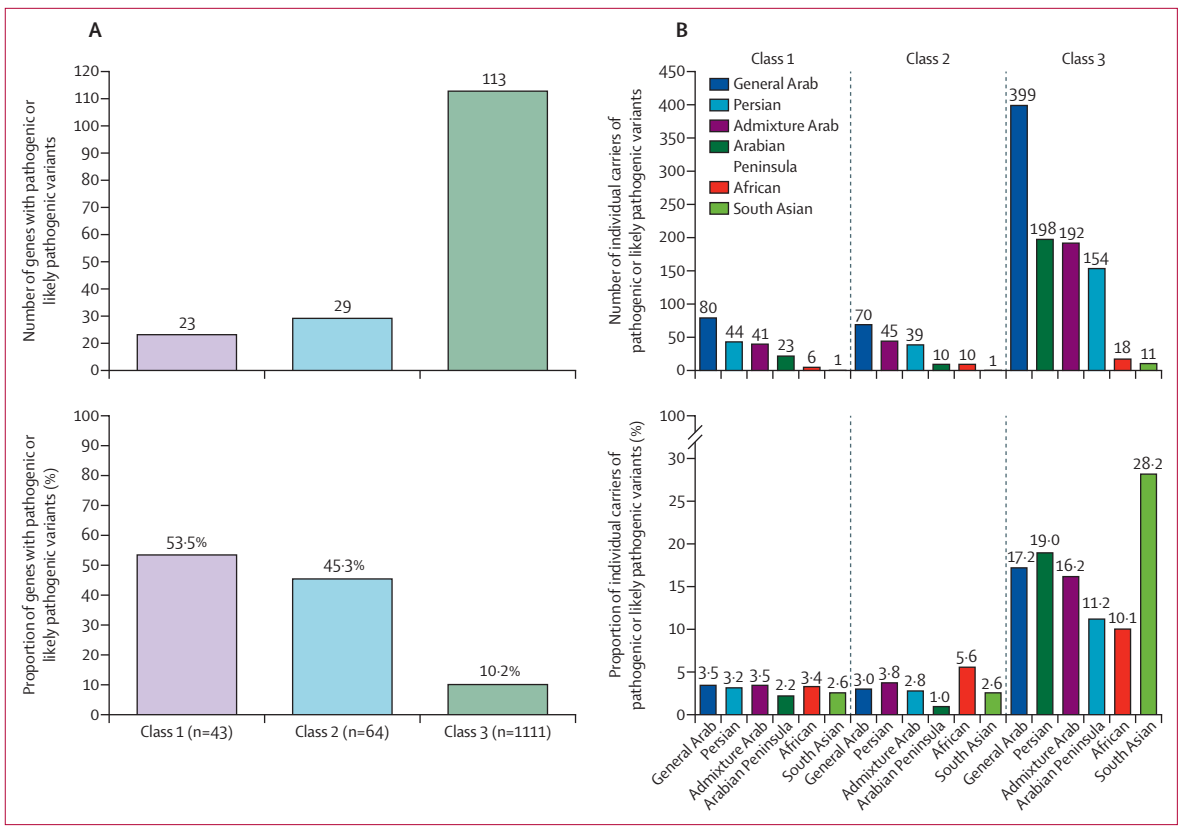

Figure 3: **Distribution of the pathogenic and likely pathogenic variants of the three cancer-gene classes in the six ancestry groups of the Qatar Genome Programme dataset**

(A) Number (upper panel) and proportion (lower panel) of genes of each cancer-gene class harbouring pathogenic or likely pathogenic variants in the Qatar Genome Programme dataset. (B) Number (upper panel) and proportion (lower panel) of pathogenic or likely pathogenic carriers in the six ancestry groups of the Qatar Genome Programme dataset.

syndrome genes, the male to female burden ratio was 1·09 (burden in women approximately 0·55% *vs* 0·57% in men).

Of the 21 individuals in the Qatar Genome Programme carrying pathogenic variants in *BRCA1*, three had a family history of cancer. One had a father with bowel cancer, and the others carried the same pathogenic variant (ie, rs80357429 [17:41223144:G:T], splice acceptor) and reported that their parents had gastric and breast cancer. Of the 18 individuals in the Qatar Genome Programme carrying pathogenic variants in *BRCA2*, three had a personal or family history of cancer. One had a mother with breast cancer, and another had a mother with gastric cancer. There was one individual with breast cancer in the Qatar Genome Programme who was carrying the variant 13:32911252:C:A, which was annotated as pathogenic only by CharGer.

EPCAM had two pathogenic variants, with one of them being present in 27 individuals and the other one being present in two individuals (figure 4C). The most frequent *EPCAM* variant (rs606231204 [2:47604159:T:TC], maximum Qatar Genome Programme minor allele frequency 0·22%) was distributed among individuals of Arabian Peninsula origin (n=17), general Arab origin (n=8), and admixture Arab origin (n=2). This variant was absent in gnomAD, version 2.1. Relatedness estimation revealed that most of the carriers were not closely related (appendix 1 p 6). Of the 27 individuals in the Qatar Genome Programme carrying *EPCAM* pathogenic variants, one individual reported a history of prostate cancer for their father.

Certain pathogenic or likely pathogenic variants of class 2 cancer genes were more frequent in the Qatar Genome Programme dataset than in public datasets (figure 4A). The most frequent pathogenic or likely pathogenic variants were found in the following class 2 cancer genes (figure 4D): *DICER1* (25 individuals) and *EXT2* (13 individuals), for which the likely pathogenic variants were identified by CharGer only; *GALNT3* (a pathogenic variant in nine individuals), *XPC* (two pathogenic variants in 12 individuals), and *PARK2* (a pathogenic variant in seven individuals). Notably, two *XPC* pathogenic variants were found almost exclusively in Qataris of general Arab origin (11 of 12 individuals). Only individuals of general Arab origin and admixture Arab origin were found to be carriers for the *PARK2* variant.

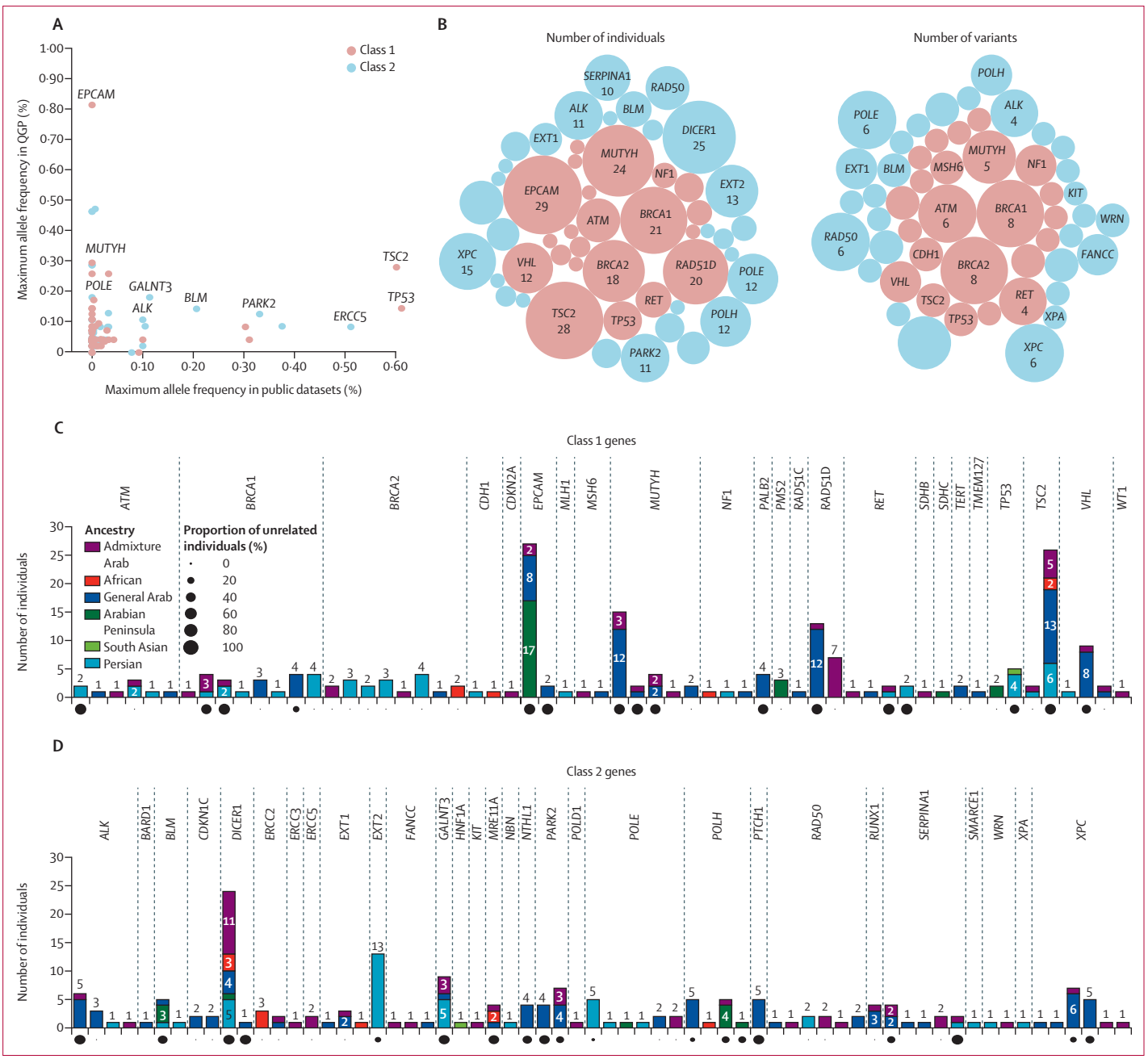

Figure 4: **Characterisation of rare pathogenic variants within key class 1 and class 2 cancer genes in the six ancestry groups of the Qatar Genome Programme dataset**
(A) Maximum of minor allele frequency (MAF) across Qatar Genome Programme (QGP) ancestry groups versus maximum of minor allele frequency across ancestry groups of pathogenic or likely pathogenic variants of key class 1 and class 2 genes in public databases. (B) Number of pathogenic or likely pathogenic variants (right) and their carriers (left) in key genes of class 1 and class 2. Only genes with more than two carriers or variants are labelled. (C) Number of carriers of pathogenic or likely pathogenic variants of key genes in the six ancestry groups in class 1. (D) Number of carriers of pathogenic or likely pathogenic variants of key genes in the six ancestry groups in class 2. The size of the filled black circles indicates the relatedness level between the carriers of the same variant.

Variants of six class 3 genes, *RAB27A, DHCR7, HMBS, MPL, MLLT1,* and *KDM2B*, were highly prevalent in Qataris and absent in public datasets (appendix 1 p 7). Two pathogenic variants of *DHCR7* were found in 55 Qataris, including 41 (74·5%) from individuals of Arabian Peninsula origin, 11 (20·0%) from those of general Arab origin, and three (5·5%) from those of admixture Arab origin (appendix 1 p 7). Two pathogenic variants in *MPL* were found exclusively in individuals of general Arab origin, Arabian Peninsula origin, and admixture Arab origin (appendix 1 p 7). One of these rare variants was found at high frequency in these groups of Qataris (109 [2·4%] of 4547 individuals from the three groups).

Of 333 pathogenic or likely pathogenic variants identified in the Qatar Genome Programme dataset, 39 were found in the Greater Middle East dataset.[22] Only two class 1 variants within *EPCAM* and *MUTYH* were obtained (appendix 2). The *EPCAM* variant (rs606231204), which was carried by 27 individuals in the Qatar Genome Programme, was only observed in individuals of Arabian Peninsula origin in the Greater Middle East dataset (minor allele frequency 0·0029). The *MUTYH* variant, rs587778541, was found in individuals of Northeast African and Turkish origin in the Greater Middle East dataset (minor allele frequency 0·0014 for those of Northeast African origin and 0·003 for those of Turkish origin; appendix 2). Six of the remaining 37 variants were in class 2 genes and 31 variants were in class 3 genes. No class 1 variants from the Qatar Genome Programme were found in the 304 non-cancer Middle Eastern samples in gnomAD (version 3).

Discussion

Our study provided in-depth screening of both common and rare cancer-gene variants in a cohort of the native Qatari population. Assessment of the heterozygosity rates in the Qatar Genome Programme cohort revealed significant variation between ancestry groups, which could reflect the differences in consanguinity rates.[14] Similarly, a high degree of heterogeneity was observed across ancestries in the polygenic risk score analyses of the most common cancers in Qatar (breast, prostate, and colorectal cancers). To avoid any potential biases in selection of polygenic risk scores, the polygenic risk scores used for the analysis were manually selected on the basis of their good performance observed in the literature and their evaluation in different ancestries along with high statistical stringency for polygenic risk score development and numbers of single-nucleotide variants included. Analysis of the selected polygenic risk scores revealed significant differences among the six ancestry-based Qatari subpopulations. However, the performance of these polygenic risk scores, which was evaluated in the TCGA dataset, might be slightly underestimated because of the comparison of the cancers included in the present analysis with all remaining cancers. This is because some single-nucleotide variants might be associated with multiple cancers. However, the impact of these overlapping associated single-nucleotide variants is expected to be small, since effect sizes are not necessarily the same across cancers. Despite recent advances that promote the use of polygenic risk scores as a potential parameter in health care and preventive strategies,[8,23] there are several limitations and debates over their clinical implementation and portability to diverse populations.[23] Polygenic risk scores, as shown in our study, can vary even within populations of a discrete geographical location such as Qatar.

Our analysis of the rare cancer-gene variants revealed several distinct characteristics of the native Qatari population. We did not assess structural cancer gene variation in this study. Current bioinformatics methods for calling structural variants are not as robust as those for calling single-nucleotide variants,[24] which means it is not possible to give reliable information for clinical interpretation without doing a subsequent experimental validation step. As for cancer-gene common variants, the distribution of the rare deleterious variants in the Qatari population was ancestry dependent. Interestingly, none of the individuals of Arabian Peninsula origin carried breast or ovarian cancer variants. Although this result suggests that this group of Qataris might have a lower genetic risk of hereditary breast or ovarian cancer, the search for *BRCA1/BRCA2* variant carriers in a larger population of Arabian Peninsula Qataris will be required to validate our finding. Some cancer variants, such as those within the *EPCAM* gene, were found to be enriched in the Qatar Genome Programme cohort and shared among unrelated individuals, supporting a potential founder-effect. Relative to their total number, class 3 genes, which are currently non-actionable cancer-related genes, contained the least number of pathogenic or likely pathogenic variants in our dataset. This is probably because they only included a small number of cancer-predisposing genes, and most of them were identified by somatic analyses, and therefore are rarely tested in a germline context. The high prevalence of certain pathogenic variants of class 3 genes causing rare recessive disorders predisposing individuals to cancer, such as variants of *DHCR7*[25,26] and *MPL*,[27] suggest that consanguinity could increase the cancer risk, particularly in children (eg, brain tumours and myeloproliferative neoplasms).

The failure to identify rare variants is crucial when it occurs in genes of known clinical importance with functionally consequential variants. This was illustrated in a participant from our cohort, who was carrying a *BRCA2* pathogenic variant detected by CharGer and not by ClinVar, and who was diagnosed with breast cancer. Although no clinical evidence was shown for this variant, our data constitute a call to action to complement the use of the ClinVar database with other computational predictors (eg, CharGer) to enhance the actionability of rare cancer-gene pathogenic or likely pathogenic variants. However, the discordance in pathogenicity prediction between different tools (eg, CharGer and InterVar) could constitute a potential challenge for their clinical use.

Incorporation of precision medicine technology, including cancer screening and genome sequencing, into the primary health-care system in Qatar has considerable potential. Although none of the participants included in our study was informed about their genetic ancestry stratification or cancer gene status, approval from the ethics committee was recently obtained, which allows sharing of genetic information and provides oncogenetics services to individuals identified as being at an increased risk of hereditary breast and ovarian cancers, as well as their families.

Other Arabian Peninsula populations are expected to have a similar genetic structure to that of Qataris. Genetic studies on these populations show that the history of human migration and geography have played an important role in shaping their genetic structure.[28,29] Therefore, it would be of interest to extend our study to other Arab populations.

There are several ethical, legal, and social issues that need to be addressed before integrating genomic-based programmes into health-care systems in Middle Eastern populations. One major challenge is the paucity of genomics studies in the Middle East. Our cohort is part of the first phase of the Qatar Genome Programme, targeting 10 000 genomes, and represents more than 10% of the total number of participants planned for inclusion in the Qatar Biobank. Proportionally to the population size (estimated to be around 300 000 individuals), the Qatar Genome Programme could be considered to be among the largest population dataset worldwide. The scarcity of individuals with cancer in our cohort prevented validation of some pathogenic variants. The absence of such data in other countries also highlights the need to generate more disease-centred data in the region, which will eventually allow large meta-analyses to study cancer and other complex traits.

Another key challenge towards the implementation of a cancer genome-based programme in the Middle East is the potential misperception of genomic studies and genetic testing by the community. In these conservative societies, individuals deal with issues including compatibility of genetic testing and religious beliefs, and a cultural fear of being genetically associated with cancer, a disease perceived as fatal, and possible stigmatisation within the community.[30] To mitigate this challenge, cross-sectoral cancer awareness programmes and public conferences addressing ontological and ethical questions raised by the genomics field have been actively underway in Qatar.

In summary, our study reports the first landscape of germline variation in the largest set of cancer-susceptibility genes from population genome sequencing of an ancestrally diverse and large Arabian cohort. The results comprise a valuable resource to capture how many cancer deleterious variants (quantitative) and for which cancers (qualitative) an individual is carrying, in ancestry-based Arab populations. With screening, prevention, and early detection at the forefront of the cancer agenda in Qatar, we propose using population genome sequencing as a means to initiate national population testing programmes to identify highly penetrant cancer gene mutation carriers. To fully deliver a precision prevention programme, more holistic large-scale studies are required, based on a combination of cancer gene mutation carriers, polygenic risk scores, clinical data, and tumour specimens.

Contributors

LC and JS conceived the study. LC, DB, YM, AR, MS, and JS designed the study. MS, MC, and DB contributed to the analysis of the TCGA dataset. RT, NS, RR, YM, and MS processed Qatar Genome Programme data. MS, NH, YM, and LC accessed and verified the data. YM, MS, KK, MS, LC, and NH contributed to the generation of the figures. YM, MS, DB, KK, MS, LC, and NH drew the figures. AR supervised actionability analysis of the datasets. LC wrote the first draft of the manuscript, with the contribution of all authors. LC, DB, AR, NH, YM, and MS did the literature search and interpreted data. YM, NH, MS did the formal analysis. LC supervised and coordinated the study. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication. All authors contributed to manuscript review and approved the final version. Members of the Qatar Genome Programme Consortium contributed to recruitment of participants, sample and data collection, phenotyping, genome sequencing, primary processing, and IT infrastructure support.

Declaration of interests

We declare no competing interests.

Data sharing

Whole-genome sequence data are not available in public repositories. They can be accessed through application to the Qatar Biobank through an established ISO-certified process by submitting a request online, subject to institutional review board approval by the Qatar Biobank.

To **submit a request** see https://www.qatarbiobank.org.qa/research/how-to-apply-new/

Acknowledgments

We are grateful to the participants who contributed to the Qatar Genome Programme study cohort. The Qatar Biobank and Qatar Genome Programme are both Research, Development & Innovation entities within the Qatar Foundation for Education, Science and Community Development. This study was funded in part by Qatar Genome Programme, Qatar National Research Fund (QNRF award PPM04-0311-200035 and NPRP12S-0317-190379). NH, JS, and MS are supported by the Biomedical Research Program at Weill Cornell Medicine-Qatar. YM is co-supported by the QNRF grant PPM1-1122-150008.

References

1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020; **70:** 7–30.

2 Brown R, Kerr K, Haoudi A, Darzi A. Tackling cancer burden in the Middle East: Qatar as an example. *Lancet Oncol* 2012; **13:** e501–08.

3 Qoronfleh MW. Pathway to excellence in cancer care: learning from Qatar's experience. *Precision Medical Sciences* 2020; **9:** 51–61.

4 Chouchane L, Boussen H, Sastry KS. Breast cancer in Arab populations: molecular characteristics and disease management implications. *Lancet Oncol* 2013; **14:** e417–24.

5 Perkins BA, Caskey CT, Brar P, et al. Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proc Natl Acad Sci USA* 2018; **115:** 3686–91.

6 Grzymski JJ, Elhanan G, Morales Rosado JA, et al. Population genetic screening efficiently identifies carriers of autosomal dominant diseases. *Nat Med* 2020; **26:** 1235–39.

7 Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* 2021; **591:** 211–19.

8 Yanes T, Young MA, Meiser B, James PA. Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field. *Breast Cancer Res* 2020; **22:** 21.

9 McGeoch L, Saunders CL, Griffin SJ, et al. Risk prediction models for colorectal cancer incorporating common genetic variants: a systematic review. *Cancer Epidemiol Biomarkers Prev* 2019; **28:** 1580–93.

10 Hughes E, Tshiaba P, Gallagher S, et al. Development and Validation of a clinical polygenic risk score to predict breast cancer risk. *JCO Precis Oncol* 2020; **4:** 4.

11 Al Thani A, Fthenou E, Paparrodopoulos S, et al. Qatar Biobank Cohort Study: study design and first results. *Am J Epidemiol* 2019; **188:** 1420–33.

12 Mbarek H, Gandhi GD, Selvaraj S, et al. Qatar Genome: insights on genomics from the Middle East. *medRxiv* 2021; published online Sept 28. https://doi.org/10.1101/2021.09.19.21263548 (preprint).

13 Razali RM, Rodriguez-Flores J, Ghorbani M, et al. Thousands of Qatari genomes inform human migration history and improve imputation of Arab haplotypes. *Nat Commun* 2021; **12:** 5929.

14 Hunter-Zinck H, Musharoff S, Salit J, et al. Population genetic structure of the people of Qatar. *Am J Hum Genet* 2010; **87:** 17–25.

15 Sayaman RW, Saad M, Thorsson V, et al. Germline genetic contribution to the immune landscape of cancer. *Immunity* 2021; **54:** 367–86.e8.

16 Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 2019; **47:** D766–73.

17 McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biol* 2016; **17:** 122.

18 Burchard EG, Ziv E, Coyle N, et al. The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med* 2003; **348:** 1170–75.

19 Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016; **44:** D862–68.

20 Scott AD, Huang KL, Weerasinghe A, et al. CharGer: clinical characterization of germline variants. *Bioinformatics* 2019; **35:** 865–67.

21 Huang KL, Mashl RJ, Wu Y, et al. Pathogenic germline variants in 10 389 adult cancers. *Cell* 2018; **173:** 355–70.e14.

22 Scott EM, Halees A, Itan Y, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat Genet* 2016; **48:** 1071–76.

23 Janssens ACJW. Validity of polygenic risk scores: are we measuring what we think we are? *Hum Mol Genet* 2019; **28:** R143–50.

24 Ebert P, Audano PA, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 2021; **372:** eabf7117.

25 Aslan A, Borcek AO, Pamukcuoglu S, Baykaner MK. Intracranial undifferentiated malign neuroglial tumor in Smith-Lemli-Opitz syndrome: a theory of a possible predisposing factor for primary brain tumors via a case report. *Childs Nerv Syst* 2017; **33:** 171–77.

26 Guizzetti M, Costa LG. Sonic hedgehog in Smith-Lemli-Opitz syndrome and tumor development. *J Pediatr Hematol Oncol* 2008; **30:** 641–42.

27 Bellanné-Chantelot C, Rabadan Moraes G, Schmaltz-Panneau B, Marty C, Vainchenker W, Plo I. Germline genetic factors in the pathogenesis of myeloproliferative neoplasms. *Blood Rev* 2020; **42:** 100710.

28 Alshamali F, Pereira L, Budowle B, Poloni ES, Currat M. Local population structure in Arabian Peninsula revealed by Y-STR diversity. *Hum Hered* 2009; **68:** 45–54.

29 Ferreira JC, Alshamali F, Montinaro F, et al. Projecting ancient ancestry in modern-day Arabians and Iranians: a key role of the past exposed Arabo-Persian Gulf on human migrations. *Genome Biol Evol* 2021; **13:** evab194.

30 El Shanti H, Chouchane L, Badii R, Gallouzi IE, Gasparini P. Genetic testing and genomic analysis: a debate on ethical, social and legal issues in the Arab world with a focus on Qatar. *J Transl Med* 2015; **13:** 358.