Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ancestry-dependent genetic structure of colorectal cancer susceptibility loci across diverse human populations

Maryam Alamil
Alfaisal University
George Kulik
Alfaisal University
Meerab Waqar
Alfaisal University
Hamza Altazi
Alfaisal University
Arwa Labkhan
Alfaisal University
Volodymyr Dvornyk
vdvornyk@alfaisal.edu

Alfaisal University

Article

Keywords: Colorectal cancer, Ancestry, GWAS, Polygenic risk score, Population structure, SNP

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10807353/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Colorectal cancer (CRC) shows marked geographic and ethnic variation in incidence and mortality, reflecting the combined effects of environmental exposures, healthcare access, and inherited susceptibility. While high-penetrance hereditary syndromes explain a minority of cases, common low-penetrance variants may contribute to inter-population differences in CRC risk. This study investigated the genetic structure and differentiation of ethnic populations at CRC susceptibility loci. We evaluated diversity parameters, allele enrichment, and unweighted population-specific polygenic risk scores (psPRS) across continental and subpopulation cohorts from the 1000 Genomes Project using 1,291 genome-wide significant CRC-associated loci. The results revealed strong, ancestry-related population stratification, with 253 loci fixed in at least one superpopulation, ranging from 16.6% in East Asians to 6.3% in Europeans. While median psPRS values hovered within a narrow, highly overlapping range across all groups, mean pairwise differences were statistically significant, reflecting the large sample size rather than major shifts in absolute cumulative genetic burden. These findings demonstrate that common CRC-susceptibility variants are non-randomly distributed, underscoring the critical need for multi-ethnic genetic studies and ancestry-aware polygenic risk modeling.

Introduction

Colorectal cancer (CRC) is a biologically heterogeneous malignancy representing one of the leading causes of cancer morbidity and mortality worldwide. In 2022, CRC accounted for approximately 9.6% of all newly diagnosed cancers globally and 9.3% of cancer deaths, placing CRC among the most important malignancies for population-level prevention, screening, and precision-risk assessment [1]. CRC is the third most often diagnosed cancer and the second deadliest cancer, with 1,926,425 cases and 904,019 deaths in 2022 worldwide. The incidence and deaths from CRC are on the rise, with 2.2 million cases and 1.1 million deaths expected in 2030 [2–4].

The etiology of CRC is multifactorial and involves the accumulation of somatic genetic and epigenetic alterations in colorectal epithelial cells. The classical adenoma–carcinoma sequence is commonly initiated by dysregulation of the APC/Wnt signaling pathway, followed by additional alterations affecting pathways such as RAS–MAPK, PI3K, TGF-β, DNA repair, and p53-mediated cell-cycle regulation [4–6]. Molecularly, CRC develops through partly overlapping routes, including chromosomal and microsatellite instability, and CpG island methylator phenotype, which contribute to differences in tumor progression, prognosis, and therapeutic response [6, 7]. Although most CRC cases are sporadic, inherited predisposition is well established: Lynch syndrome is caused mainly by germline pathogenic variants in DNA mismatch repair genes, including *MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM*, whereas familial adenomatous polyposis is caused by pathogenic *APC* variants [8–10]. Other hereditary or familial syndromes involve genes such as *BMPR1A*, *MUTYH*, *NTHL1*, *POLE*, *POLD1*, *PTEN*, *SMAD4*, and *STK11*, demonstrating the broad range of genetic alterations underlying CRC susceptibility [8, 10, 11].

In parallel with high-penetrance hereditary syndromes, common low-penetrance variants contribute substantially to inter-individual and inter-population variation in CRC risk. Genome-wide association studies (GWAS) have identified several hundred common variants independently associated with CRC risk, most of which lie in non-coding or intergenic regions and are presumed to influence regulatory mechanisms rather than the protein-coding sequence directly [12].

Non-genetic factors play an important role in CRC etiology, and lifestyle and socio-economic differences may explain much of the global and ethnic variation in disease burden. Diets high in red and processed meat, alcohol consumption, adiposity, physical inactivity, and smoking are associated with increased CRC risk, whereas dietary patterns rich in whole grains, fiber, and plant-based foods and adherence to cancer-prevention lifestyle decrease the risk [13–17]. A recent systematic review of dietary-lifestyle patterns concluded that greater adherence to WCRF/AICR recommendations was supported by “strong–probable” evidence for lower CRC risk, reinforcing the importance of modifiable exposures in primary prevention [16]. Such exposures vary across ethnic groups due to cultural dietary practices, urbanization, migration, occupational activities, the built environment, income, education, and healthcare access. Therefore, ethnic differentiation at CRC candidate loci should be interpreted within an exposomic and social context rather than as evidence of genetic determinism.

The global burden of CRC is not evenly distributed: incidence is generally highest in industrialized regions, including Europe, North America, Australia, and New Zealand, whereas historically lower rates have been reported in parts of Africa and South-Central Asia [1, 6]. Ethnic disparities in CRC prevalence and outcomes are especially evident within multiethnic countries. In the United States, Black people have consistently shown higher CRC incidence and poorer survival than White people, Asian and Hispanic groups, although substantial heterogeneity exists within these broad categories [7, 18]. The largest share of observed survival disparity was attributed to neighborhood socioeconomic status, followed by tumor molecular characteristics and sidedness [7]. These observations emphasize that ethnic differences in CRC burden cannot be attributed solely to genetics; rather, they emerge from interactions among genetic ancestry, tumor biology, screening access, socioeconomic position, healthcare quality, comorbidities, and treatment availability.

Despite the significant role of non-genetic factors, population differences in mutations associated with CRC susceptibility may contribute to observed differences in CRC incidence and mortality between ethnic groups.

Thus, cancer registry data in the USA show substantial differences in age-standardized mortality rates per 100,000 between ethnic groups: most differences are found between non-Hispanic Asian/Pacific Islanders, Non-Hispanic Black people, and Non-Hispanic White people [19]. Likewise, a recent study showed higher CRC mortality for non-Hispanic White people as compared to Asian Indians [20].

Similar differences between ethnic groups are observed in the UK, where the age-adjusted incidence ratio for CRC in British Indians compared with British White individuals was 0.46 [21]. Another UK study

of CRC incidence in Wolverhampton showed that the odds ratio for colorectal cancer in South Asians was 0.2 ($p < 0.001$) compared to non-South Asians (77.79% white British) [22].

The problem of ancestry-dependent disparities in disease prevalence has been addressed quite extensively (e.g., [23–25]). Between-population differences in genetic structure were suggested among the factors contributing to the above disparities [26–30].

Using data from the 1000 Genomes Project [31], we examined the genetic structure and macro-level population stratification of diverse human cohorts across 1,291 CRC susceptibility loci. To bypass Eurocentric discovery weights and avoid clinical skewing, we apply an unweighted multi-locus structural approach. Our goal is to evaluate whether known common disease variants mirror deep demographic evolutionary history or demonstrate specific frequency shifts, highlighting the structural parameters that must be accounted for when modeling genetic risk across diverse human ancestries.

Materials and Methods

The data

We analyzed the genetic structure of the populations included in the 1000 Genomes Project, i.e., five major superpopulations and 26 subpopulations (Supplementary Table 1). Candidate loci for CRC with genome-wide significance were extracted from the GWAS catalog (https://www.ebi.ac.uk/gwas/, accessed on 25 March 2026). The results of GWAS for the following traits were screened: “colorectal cancer”, “metastatic colorectal cancer”, and “colorectal carcinoma”. To avoid bias, loci with missing genotypes and/or without dbSNP ID were excluded from the analysis. The final set included 1291 loci. Individual genotypes were retrieved from the 1000 Genomes Phase 3 Project database [32] at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ (accessed on 29 March 2026).

Analysis of population genetic structure and allele frequencies

All loci were checked for correspondence to the Hardy-Weinberg equilibrium (HWE). The standard parameters of population genetic structure were estimated: allele frequencies, observed (H_o) and expected (H_e) heterozygosity [24], and Wright’s fixation index (F, inbreeding coefficient) [33]. We used Φ_{PT} from AMOVA (analysis of molecular variance) [34, 35], Nei’s genetic distances [36], and the log-likelihood ratio (G) based exact test [37] to estimate interpopulation differentiation. A pairwise interpopulation Φ_{PT} matrix was used for the Principal Coordinate Analysis (PCoA). All analyses were performed using GenAlEx v. 6.5 [35] and GENEPOP v. 4.8.3 [38].

We constructed a heatmap to visualize the relative enrichment/depletion of risk alleles in each population with respect to global allele frequencies across loci. Hypergeometric distribution–based p-values were transformed to a log_{10} scale and adjusted accordingly to indicate allele enrichment

(positive values) or depletion (negative values). The populations and alleles were clustered according to their average correlation structure.

To compute the individual composite polygenic risk score (PRS) for CRC, we followed the approach described by Choi *et al*. and implemented in the PRSice-2 v. 2.3.5 software [39]:

$$PRS = \sum_{i=1}^{N} (\beta_i \times SNP_i)$$

,

where β_i denotes the effect size of the *i*-th SNP, SNP_i is the allele dosage (0, 1, or 2 under an additive model) for that SNP in the individual, and *N* is the total number of SNPs shared between the base and target datasets. We explicitly omitted population-specific effect size weights (β) from our cumulative score computation to mitigate the well-documented transferability decay of Eurocentric GWAS discoveries across diverse ancestral backgrounds [40]. Since applying European-derived discovery weights to non-European cohorts systematically skews clinical risk prediction, the unweighted approach is utilized here strictly as a structural metric. This score serves exclusively as a proxy for mapping multi-locus allele frequency stratification across geography and evolutionary history, rather than as a predictive tool for absolute clinical risk or diagnostic penetrance. Therefore, the final formula was

$$PRS = \sum_{i=1}^{N} SNP_i$$

.

While the structural utility of this metric relies on the partial cross-population portability of primary GWAS index signals, it was acknowledged that shifting linkage disequilibrium blocks and variant frequencies across diverse ancestral backgrounds remain major confounding factors that limit the direct translation of these associations [41]. Population-specific PRS (psPRS) values were computed as the mean PRS for each population.

The interpopulation differences in psPRS were assessed by a non-parametric statistical approach. First, the Kruskal–Wallis rank-sum test was applied to check for possible significant differences between populations, assuming a non-normal distribution. If the test yielded a statistically significant *p*-value (< 0.05), Dunn's test was applied to identify the differences between all population pairs. The Dunn's test *p*-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) correction. All statistical analyses were performed in the R environment using the FSA package v. 0.10.1 [42].

Results

Allelic structure of the populations

The populations differed quite significantly in their allelic structure. Out of 1291 loci analyzed, 253 were fixed in at least one superpopulation (Supplementary Table 2). However, the number of monomorphic loci was distributed unevenly across the superpopulations. The largest proportion of monomorphic loci, 16.6%, was observed in EAS, followed by 10.2% in SAS, 10.1% in AFR, 6.9% in AMR, and 6.3% in EUR. At the subpopulation level, the percentage of polymorphic loci within the same ancestry group did not vary substantially, except for two African diaspora populations in North America, ACB and ACW, which had, on average, about 10% more polymorphic loci than the other subpopulations of this ancestry.

In total, 330 loci manifested a departure from the HWE in any one of the superpopulations, whereas 60 more loci – in two or more superpopulations. Out of these, ten loci (rs5002178, rs12700664, rs199562, rs34246098, rs35409710, rs75679567, rs9271695, rs9271770, rs9273736, rs9952412) significantly deviated in three or more superpopulations. The proportion of loci deviating from HWE in subpopulations ranged from 3 to 7%.

Heatmaps of allele frequencies across subpopulations revealed pronounced ancestry-related structure among CRC–associated variants (Fig. 1A). Populations clustered primarily by shared genetic ancestry, with clear separation among African, European, East Asian, South Asian, and admixed American groups. Closely related subpopulations within continental groupings showed similar enrichment and depletion profiles, whereas substantial divergence was observed between populations of different continental origin. Large blocks of loci exhibited coordinated enrichment or depletion within specific subpopulations, indicating that CRC-associated alleles are unevenly distributed worldwide.

Ancestry-dependent between-population differences were more pronounced among the continental superpopulations (Fig. 1B). European and East Asian populations displayed largely distinct and, in multiple variant clusters, reciprocal enrichment profiles. South Asian and admixed American populations showed intermediate patterns. African superpopulation formed the most divergent cluster, characterized by strong bidirectional deviation across variant sets, which may partly reflect the fact that most currently known CRC susceptibility loci were discovered in predominantly European GWAS cohorts.

The broad white band observed across populations corresponds to a subset of CRC susceptibility loci showing little inter-population variation in effect allele frequencies. These loci appear neither enriched nor depleted in any specific ancestry group, suggesting that they are unlikely to account for major between-population differences in CRC prevalence. Rather, population differences in disease burden are more likely to be influenced by the subset of loci showing marked frequency contrasts among populations.

Together, these results demonstrate that common CRC-associated variants exhibit strong population stratification at both fine-scale and continental levels.

Intra- and interpopulation genetic diversity

The genetic diversity parameters for the five superpopulations are given in Table 1. The highest diversity among the superpopulations was observed in EUR, and the lowest in AFR. The subpopulations within the same continental group showed similar diversity levels, though variation was higher among the African and American subpopulations. All populations were characterized by very low inbreeding levels (Supplementary Table 3).

Table 1
The parameters of genetic variation at CRC risk loci across superpopulations.

Populations	H_o	H_e	*I*	*F*
AFR	0.230 ± 0.005	0.232 ± 0.005	0.357 ± 0.007	0.009 ± 0.001
AMR	0.272 ± 0.005	0.281 ± 0.005	0.425 ± 0.007	0.025 ± 0.002
EAS	0.246 ± 0.005	0.249 ± 0.005	0.375 ± 0.007	0.011 ± 0.001
EUR	0.285 ± 0.005	0.287 ± 0.005	0.432 ± 0.007	0.008 ± 0.001
SAS	0.274 ± 0.005	0.278 ± 0.005	0.419 ± 0.007	0.016 ± 0.001

H_o, observed heterozygosity; H_e, expected heterozygosity; *I*, Shannon's information index; *F*, fixation index.

Both the AMOVA and principal coordinates analysis (PCoA) revealed significant population differentiation at the CRC candidate loci (Table 2, Supplementary Table 4). In the PCoA plot (Fig. 2B), the European, South Asian, and East Asian subpopulations form tight, well-defined clusters, indicating low genetic variation within continents and high similarity among the groups. The East Asian and African populations are clearly separated from the others. The American admixed subpopulations (PUR, CLM, MXL, PEL) sit between Europeans and South Asians, reflecting different proportions of European, Native American, and (in some cases) African ancestry. Likewise, American and South Asian subpopulations are located between EUR and EAS/AFR, showing varying levels of admixture typical of these groups.

Table 2
A matrix of the pairwise Φ_{PT} values and Nei's genetic distances between the superpopulations.

AFR	**AFR**	**AMR**	**EAS**	**EUR**	**SAS**
		0.048	**0.070**	**0.054**	**0.034**
AMR	0.217		0.027	0.011	0.015
EAS	0.296	0.127		0.043	0.028
EUR	0.234	0.047	0.185		0.017
SAS	0.201	0.065	0.130	0.075	

The Φ_{PT} values are given below the diagonal; the Nei's genetic distances are given above the diagonal. All Φ_{PT} values are significant at $p \leq 0.001$.

The significant interpopulation differentiation was further supported by the results of the pairwise exact *G*-test for genic and genotypic differentiation [37]. The analysis across all 1291 loci yielded highly significant differentiation ($p < < 10^{-4}$) between all pairs of superpopulations (the results not shown). Similarly, significant between-population differentiation was observed at the subpopulation level: only three pairwise comparisons, all within the same ancestries, CEU – GBR, ESN – YRI, and ITU – STU, yielded non-significant *p*-values (Supplementary Table 5).

Composite genetic risk scores at the CRC candidate loci

The median psPRS of all super-populations varied modestly around ~ 0.05, thus suggesting no extreme divergence in polygenic burden across continents. However, despite small differences in psPRS absolute values among the superpopulations, all pairwise differences (except two, EAS – EUR and AMR – SAS) were statistically significant (Table 3). The mean unweighted composite genetic risk score for CRC across all studied populations was 0.049 ± 0.005, with the lowest in admixed Americans (0.048 ± 0.004) and the highest in Africans (0.051 ± 0.006; Fig. 3A).

The scores across subpopulations showed greater heterogeneity, with generally smaller between-population differences (Fig. 3B). The largest differences were observed between the African and non-African subpopulations. However, some population pairs of the same ancestry also manifested quite significant differences, e.g., CDX – CHS and CDX – JPT (Table S6).

Table 3
The *p*-values of the psPRS pairwise comparisons across superpopulations using Dunn's test with Benjamini–Hochberg correction.

Population pair	Z	*P*unadj	*P*adj	Significance
AFR - AMR	9.345	9.17×10^{-21}	9.17×10^{-20}	***
AFR - EAS	4.715	2.41×10^{-6}	4.02×10^{-6}	***
AMR - EAS	-4.884	1.04×10^{-6}	2.08×10^{-6}	***
AFR - EUR	2.984	0.0028	0.0036	**
AMR - EUR	-6.347	2.19×10^{-10}	7.30×10^{-10}	***
EAS - EUR	-1.623	0.1046	0.1162	ns
AFR - SAS	9.190	3.94×10^{-20}	1.97×10^{-19}	***
AMR - SAS	-1.017	0.3092	0.3092	ns
EAS - SAS	4.242	2.21×10^{-5}	3.16×10^{-5}	***
EUR - SAS	5.851	4.89×10^{-9}	1.22×10^{-8}	***
** $p < 0.01$, *** $p < 0.001$				

Although most pairwise differences in psPRS values were statistically significant, the absolute differences between superpopulations were small. Mean psPRS values ranged from 0.048 in AMR to 0.051 in AFR, corresponding to an approximately 6% relative difference across the observed range. Therefore, the statistical significance of the comparisons largely reflects the large number of analyzed loci and the sample size rather than major shifts in cumulative genetic burden. The distributions of scores across populations were highly overlapping (Fig. 3), suggesting that most of individual-level variation in polygenic susceptibility occurs within rather than between populations.

Discussion

Consistent significant deviations from HWE at some loci across the populations

Among the loci consistently deviating from HWE across the superpopulations, several (rs9271695, rs9271770, rs9273736, rs35409710, and rs5002178) are located in the *HLA/MHC* region on chromosome 6. This is a highly polymorphic gene cluster, essential for immune system function, and exhibiting high linkage disequilibrium [43, 44]. There is evidence that this region undergoes strong balancing selection [45, 46]. All these factors apparently contribute to the observed deviations from HWE at these loci across population groups.

The biological significance of the remaining loci is less well characterized. Their repeated departure from HWE may reflect selection, local linkage patterns, population structure, or other evolutionary processes, although additional analyses would be required to distinguish among these possibilities.

The PCoA results reveal a strong global stratification of genetic variation at both the sub- and super-population levels. Each continental group forms a distinct and coherent cluster, while admixed populations occupy predictable intermediate positions. African populations appear to be the most genetically distinct, reflecting deeper population history. Collectively, the diagrams are consistent with the well-established structure of human populations and indicate that ancestry is apparently a major determinant of allele frequency distributions at the loci analyzed. This suggests that CRC risk alleles segregate within structured populations rather than a panmictic global pool and therefore disease-associated variation should be interpreted within a population-genetic framework.

Contrary to genome-wide expectations, African populations did not exhibit the highest genetic diversity at the analyzed CRC-candidate loci. Instead, several European, South Asian, and admixed American populations showed higher diversity and heterozygosity (Supplementary Table 4). The reduced diversity observed in African populations at CRC-associated loci may reflect ascertainment bias arising from the predominance of European-ancestry GWAS, differences in linkage disequilibrium structure, or locus-specific evolutionary processes. Distinguishing among these explanations will require future analyses incorporating ancestry-diverse association studies and formal tests of selection.

The observed genetic patterns show only partial concordance with global CRC epidemiological data. Public cancer surveillance data consistently report the highest CRC incidence in high-income regions, particularly Europe, North America, and parts of East Asia—and the lowest in most African regions, with age-standardized incidence rates several-fold higher in Europe than in sub-Saharan Africa. Regions with higher Human Development Index (HDI) have significantly higher CRC incidence, due to differences in screening methods, longevity, diets, obesity, and lifestyle factors [2, 47]. Despite this, the variations in unweighted polygenic risk scores observed here were minor and did not mirror these global incidence patterns. While African cohorts statistically exhibited the highest mean psPRS and admixed Americans the lowest within our specific locus set, the absolute differences across the entire range were minimal (approximately 6%), highlighting that these scores map multi-locus allele frequency stratification rather than clinical disease risk. Moreover, several African subpopulations displayed median scores comparable to or exceeding those of European and East Asian populations. This apparent mismatch between genetic susceptibility and disease incidence suggests that currently identified CRC susceptibility loci capture only part of the underlying genetic architecture of CRC. It further highlights the dominant role of environmental, dietary, demographic, healthcare-access, and socioeconomic factors in shaping global patterns of CRC burden [2, 47, 48]. Therefore, ancestry-specific genetic architecture should be interpreted as just one component of CRC susceptibility rather than a direct explanation for regional differences in CRC incidence.

Finally, the distribution of polygenic risk scores and weak but persistent deviations from HWE across populations illustrate the complexity of transferring genetic risk models across ancestries [49]. PRS distributions showed generally similar central tendencies across superpopulations but differed in variance and tail behavior, patterns known to arise from ancestry-specific allele frequencies, linkage disequilibrium differences, and Eurocentric GWAS discovery bias. Furthermore, HWE deviations at CRC-candidate loci across multiple populations may also suggest selection operating on or near functional variants rather than technical artifacts (see above). Therefore, differences in genetic diversity or PRS distributions should not be interpreted as direct indicators for disease prevalence; epidemiological data clearly demonstrate that CRC prevalence is heavily influenced by non-genetic factors [50, 51]. Rather, our findings reiterate that CRC susceptibility results from a complex interplay between population demographic history, locus-specific evolutionary forces, and various non-genetic factors. They emphasize the need for ancestry-conscious approaches when interpreting genetic risk across diverse human populations.

Differentiation at CRC candidate loci is above the level expected from genome-wide variation

An important observation is that the magnitude of differentiation at CRC susceptibility loci exceeds the level expected from genome-wide population structure. In human populations, genome-wide differentiation between major continental groups is typically modest, with autosomal F_{ST} values generally ranging from approximately 0.05 to 0.15 [52]. In contrast, pairwise differentiation at CRC susceptibility loci reached Φ_{PT} values of 0.296 (AFR-EAS), 0.234 (AFR-EUR), and 0.185 (EAS-EUR). Although Φ_{PT} and F_{ST} are mathematically distinct metrics with different baseline ranges and cannot be directly compared, these prominent values tentatively indicate a notable encapsulation of ancestry-informative variation among the analyzed CRC-associated loci.

Importantly, the topology of between-population relationships remained consistent with established human demographic history. AFR-EAS represented the most differentiated population pair, EUR-SAS the least differentiated major continental pair, and AMR occupied an intermediate position reflecting admixture. Thus, CRC susceptibility loci retain the expected demographic signal while exhibiting elevated between-population allele-frequency divergence. This pattern suggests enrichment of ancestry-informative variation among CRC-associated loci and emphasizes that population structure should be taken into account when transferring genetic risk models across ancestry groups.

Ancestry-dependent genetic architecture of colorectal cancer susceptibility

The observed population differentiation suggests that the genetic architecture of CRC susceptibility related to common variants is strongly shaped by ancestry. Large-scale GWAS have identified hundreds of loci contributing to CRC risk, most of which produce small individual effects but collectively account

for substantial risk [53, 54]. Many of these loci show significant differences in allele frequencies across populations, consistent with the structured patterns observed here.

The loci represented in these analyses are primarily noncoding regulatory variants influencing central pathways of colorectal homeostasis, including Wnt/β-catenin signaling, TGF-β/SMAD signaling, immune regulation, and epithelial integrity [53, 55]. Such variants show substantial population differentiation, although the evolutionary mechanisms underlying these patterns remain to be determined.

The ancestry-dependent distribution of CRC susceptibility loci is consistent with patterns reported for several other complex human diseases, including cancers, with documented between-population differences in genetic risk and disease prevalence. In breast cancer, genetic susceptibility models incorporating ancestry-specific allele frequencies, effect sizes, and population-specific variants were shown to influence risk estimates [56]. Likewise, studies of prostate cancer and glioblastoma multiforme in diverse cohorts have shown that genetic susceptibility varies across ancestry groups [57, 58]. A worldwide analysis of 14 common disorders showed that average GWAS-based genetic risk for several diseases differs among continental groups and correlates significantly with population prevalence, suggesting that ancestry-related genetic structure likely contributes to the observed discrepancies [59]. Multi-ancestry studies of type 2 diabetes have similarly demonstrated that combining GWAS data from Europeans, Africans, and East Asians captures both shared and population-specific components of disease susceptibility, reflecting ancestry-related differences in allele frequencies and linkage disequilibrium structure [60].

Implications for polygenic risk scores and health equity

The strong effect of ancestry on allele frequencies of CRC candidate loci has important implications. Several studies have demonstrated that CRC PRSs derived largely from European cohorts have limited predictive accuracy in non-European populations, particularly those of African ancestry [52, 61]. The divergent patterns observed here underscore the need for ancestry-aware calibration and multi-ancestry training datasets when constructing tools for genetic risk prediction.

Otherwise, clinical implementation of PRSs may yield biased estimates of existing disparities in CRC prevention and early detection. Expanding population diversity in GWAS and fine-mapping studies is therefore essential for both locus discovery and the unbiased translation of genetic findings into public health and clinical practice.

Common versus hereditary colorectal cancer genetics

The between-population differences observed in this study contrast with the genetic architecture of hereditary CRC syndromes, such as Lynch syndrome or familial adenomatous polyposis. These diseases are driven by rare, high-penetrance mutations, which are under strong negative selection and show weak population specificity [62]. Our results support the polygenic basis of sporadic CRC, where many common variants with small individual effects collectively shape risk at the population level. However,

while common variants may contribute to population-level differences in CRC susceptibility, the impact of rare high-penetrance mutations is likely more important at the individual and familial level.

Limitations

Several structural limitations must be considered when interpreting these findings. First, the candidate variants analyzed were primarily discovered in cohorts of predominantly European ancestry within the GWAS Catalog, introducing an inherent ascertainment bias that impacts the observed diversity parameters in non-European groups—particularly the African superpopulation. Second, while our unweighted polygenic approach successfully captures broad, macro-level population stratification at disease-associated loci, it lacks the resolution to account for linkage disequilibrium structural differences or population-specific effect sizes. Consequently, these genetic risk score discrepancies do not directly mirror real-world epidemiological variations in CRC incidence.

Conclusions

Our results demonstrate that common variants associated with CRC are distributed non-randomly across human populations. This reflects both deep evolutionary history and more recent demographic events. The ancestry-dependent population genetic structure provides insights into the genetic component of CRC susceptibility and may contribute to understanding inter-population differences in disease risk when considered alongside environmental and socioeconomic factors. Our results further emphasize the need for ancestry-aware and trans-ancestry polygenic risk models to ensure equitable implementation of precision prevention strategies.

Abbreviations

AICR
American Institute for Cancer Research
AMOVA
Analysis of Molecular Variance
APC
Adenomatous Polyposis Coli
BMPR1A
Bone Morphogenetic Protein Receptor 1A
CRC
Colorectal cancer
EPCAM
Epithelial Cell Adhesion Molecule
GWAS
Genome-Wide Association Study
HDI
Human Development Index

HLA
Human Leukocyte Antigen
HWE
Hardy-Weinberg Equilibrium
MAPK
Mitogen-Activated Protein Kinase
MHC
Major Histocompatibility Complex
MLH1
MutL Homolog 1
MSH2
MutS Homolog 2
MSH6
MutS Homolog 6
MUTYH
mutY DNA Glycosylase
NTHL1
Nth Endonuclease III-Like 1
PCoA
Principal Coordinates Analysis
PI3K
Phosphoinositide 3-Kinase
PMS2
Postmeiotic Segregation Increased 2
POLD1
DNA Polymerase Delta 1
POLE
DNA Polymerase Epsilon
PRS
Polygenic Risk Score
psPRS
Population-Specific Polygenic Risk Score
PTEN
Phosphatase and Tensin Homolog
RAS
Rat Sarcoma
SMAD
Sma and Mad Homolog
STK11
Serine/Threonine Kinase 11

TGF-β
Transforming Growth Factor-Beta
WCRF
World Cancer Research Fund
Wnt
Wingless-related integration site

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

The authors received no funding for this study.

Competing interests

The authors declare that they have no competing interests.

Data availability

No new data were created or analyzed in this study. The original individual genotype data for each analyzed locus are freely available from the 1000 Genomes Project at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ (accessed on March 5, 2025).

Author Contribution

M.A.: investigation, methodology, resources, formal analysis, writing—original draft. writing—review and editing. G.K.: investigation, methodology, resources, formal analysis, writing—original draft, writing—review and editing. M.W., H.A., A.L.: resources, data curation. V.D.: conceptualization, formal analysis, methodology, validation, writing—original draft, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Data Availability

No new data were created or analyzed in this study. The original individual genotype data for each analyzed locus are freely available from the 1000 Genomes Project at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ (accessed on March 5, 2025).

References

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263. 10.3322/caac.21834 (2024).
2. Wu, S., Zhang, Y., Lin, Z. & Wei, M. Global burden of colorectal cancer in 2022 and projections to 2050: incidence and mortality estimates from GLOBOCAN. *BMC Cancer*. **25**, 1770. 10.1186/s12885-025-15138-0 (2025).
3. De Rosa, M. et al. Genetics, diagnosis and management of colorectal cancer (Review). *Oncol. Rep.* **34**, 1087–1096. 10.3892/or.2015.4108 (2015).
4. Li, J., Ma, X., Chakravarti, D., Shalapour, S. & DePinho, R. A. Genetic and biological hallmarks of colorectal cancer. *Genes Dev.* **35**, 787–820. 10.1101/gad.348226.120 (2021).
5. Sawicki, T. et al. A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. *Cancers* 13, (2025). 10.3390/cancers13092025 (2021).
6. Hull, R., Francies, F. Z., Oyomno, M. & Dlamini, Z. Colorectal cancer genetics, incidence and risk factors: In search for targeted therapies. *Cancer Manag Res.* **12**, 9869–9882. 10.2147/CMAR.S251223 (2020).
7. Yousef, M., Yousef, A. & Chowdhury, S. Molecular, socioeconomic, and clinical factors affecting racial and ethnic disparities in colorectal cancer survival. *JAMA Oncol.* **10**, 1519–1529. 10.1001/jamaoncol.2024.3666 (2024).
8. Chen, L., Ye, L. & Hu, B. Hereditary colorectal cancer syndromes: Molecular genetics and precision medicine. *Biomedicines* **10**, 3207. 10.3390/biomedicines10123207 (2022).
9. Li, X., Liu, G. & Wu, W. Recent advances in Lynch syndrome. *Exp. Hematol. Oncol.* **10** 10.1186/s40164-021-00231-4 (2021).
10. Talseth-Palmer, B. A. The genetic basis of colonic adenomatous polyposis syndromes. *Hered Cancer Clin. Pract.* **15** 10.1186/s13053-017-0065-x (2017).
11. Kim, J. C. & Bodmer, W. F. Genotypic and phenotypic characteristics of hereditary colorectal cancer. *Ann. Coloproctol*. **37**, 368–381. 10.3393/ac.2021.00878.0125 (2021).
12. Chen, Z. et al. Fine-mapping analysis including over 254,000 East Asian and European descendants identifies 136 putative colorectal cancer susceptibility genes. *Nat. Commun.* **15**, 3557. 10.1038/s41467-024-47399-x (2024).
13. Veettil, S. K. et al. Role of Diet in Colorectal Cancer Incidence: Umbrella Review of Meta-analyses of Prospective Observational Studies. *JAMA Netw. Open.* **4**, e2037341. 10.1001/jamanetworkopen.2020.37341 (2021).

14. Vallis, J. & Wang, P. P. in *Gastrointestinal Cancers* (ed J. A. Morgado-Diaz) (2022).
15. Clinton, S. K., Giovannucci, E. L., Hursting, S. D. & The World Cancer Research Fund/American Institute for Cancer Research Third Expert Report on Diet. Nutrition, Physical Activity, and Cancer: Impact and Future Directions. *J. Nutr.* **150**, 663–671. 10.1093/jn/nxz268 (2020).
16. Chu, A. H. Y. et al. Dietary-lifestyle patterns and colorectal cancer risk: Global Cancer Update Programme systematic literature review. *Am. J. Clin. Nutr.* **121**, 986–998 (2025).
17. Chu, A. H. Y. et al. Dietary patterns and colorectal cancer risk: Global Cancer Update Programme systematic literature review. *Am. J. Clin. Nutr.* **121**, 999–1016 (2025).
18. Riviere, P. et al. Racial disparities in colorectal cancer outcomes and access to care: a multi-cohort analysis. *Front. Pub Health*. **12**, 1414361. 10.3389/fpubh.2024.1414361 (2024).
19. Carethers, J. M. Racial and ethnic disparities in colorectal cancer incidence and mortality. *Adv. Cancer Res.* **151**, 197–229. 10.1016/bs.acr.2021.02.007 (2021).
20. Thakkar, Z. et al. Disaggregated colorectal cancer mortality among Asian American subgroups between 2005 and 2020. *Cancer Epidemiol. Biomarkers Prev.* **34**, 1134–1140. 10.1158/1055-9965.EPI-24-1688 (2025).
21. Ali, R., Barnes, I., Kan, S. W. & Beral, V. Cancer incidence in British Indians and British whites in Leicester, 2001–2006. *Br. J. Cancer*. **103**, 143–148. 10.1038/sj.bjc.6605744 (2010).
22. Hebbar, S., Fuggle, W. J., Nevill, A. M. & Veitch, A. M. Colorectal cancer incidence and trend in UK South Asians: a 20-year study. *Colorectal Dis.* **14**, e319–322. 10.1111/j.1463-1318.2012.02960.x (2012).
23. Global Burden of Disease. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **396**, 1204–1222. 10.1016/s0140-6736(20)30925-9 (2020).
24. Du, Y. R. et al. Ethnic disparities in prevalence of chronic non-communicable diseases and its multimorbidity among older adults in rural southwest China. *BMC Public. Health*. **23**, 1217. 10.1186/s12889-023-16161-1 (2023).
25. Langer-Gould, A. M., Gonzales, E. G., Smith, J. B., Li, B. H. & Nelson, L. M. Racial and ethnic disparities in multiple sclerosis prevalence. *Neurology* **98**, e1818–e1827. 10.1212/wnl.0000000000200151 (2022).
26. Mavrych, V., Alamil, M., Bolgova, O. & Dvornyk, V. Genetic structure of different ethnic populations at the frontotemporal dementia risk loci. *PLoS One*. **20**, e0329809. 10.1371/journal.pone.0329809 (2025).
27. Dvornyk, V. Significant interpopulation differentiation at candidate loci may underlie ethnic disparities in the prevalence of uterine fibroids. *J. Genet.* **100** 10.1007/s12041-021-01342-x (2021).
28. Dvornyk, V. et al. Differentiation of Caucasians and Chinese at bone mass candidate genes: implication for ethnic difference of bone mass. *Ann. Hum. Genet.* **67**, 216–227. 10.1046/j.1469-1809.2003.00037.x (2003).

29. Mao, L., Fang, Y., Campbell, M. & Southerland, W. M. Population differentiation in allele frequencies of obesity-associated SNPs. *BMC Genom.* **18**, 861. 10.1186/s12864-017-4262-9 (2017).
30. Yoon, B. W., Shin, H. T. & Seo, J. H. Risk allele frequency analysis and risk prediction of single-nucleotide polymorphisms for prostate cancer. *Genes* **13** 10.3390/genes13112039 (2022).
31. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74. 10.1038/nature15393 (2015).
32. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073. 10.1038/nature09534 (2010).
33. Weir, B. S. & Cockerham, C. C. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370. 10.1111/j.1558-5646.1984.tb05657.x (1984).
34. Meirmans, P. G. Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* **60**, 2399–2402 (2006).
35. Peakall, R. & Smouse, P. E. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. *Bioinformatics* **28**, 2537–2539. 10.1093/bioinformatics/bts460 (2012).
36. Nei, M. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583–590. 10.1093/genetics/89.3.583 (1978).
37. Goudet, J., Raymond, M., de Meeus, T. & Rousset, F. Testing differentiation in diploid populations. *Genetics* **144**, 1933–1940. 10.1093/genetics/144.4.1933 (1996).
38. Rousset, F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* **8**, 103–106. 10.1111/j.1471-8286.2007.01931.x (2008).
39. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8** 10.1093/gigascience/giz082 (2019).
40. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649. 10.1016/j.ajhg.2017.03.004 (2017).
41. Carlson, C. S. et al. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* **11**, e1001661. 10.1371/journal.pbio.1001661 (2013).
42. FSA: Simple Fisheries Stock Assessment Methods. (2026).
43. Wassenaar, T. M., Harville, T., Chastain, J., Wanchai, V. & Ussery, D. W. DNA structural features and variability of complete *MHC* locus sequences. *Front. Bioinf.* **4** 10.3389/fbinf.2024.1392613 (2024).
44. Lim, J., Bae, S. C., Kim, K. & Understanding *HLA* associations from SNP summary association statistics. *Sci. Rep.* 9, 1337, (2019). 10.1038/s41598-018-37840-9
45. Slatkin, M. Joint estimation of selection intensity and mutation rate under balancing selection with applications to HLA. *Genetics* **221** 10.1093/genetics/iyac058 (2022).
46. Black, F. L. & Hedrick, P. W. Strong balancing selection at *HLA* loci: evidence from segregation in South Amerindian families. *Proc. Natl Acad. Sci. U.S.A.* 94, 12452–12456 (1997).

47. Morgan, E. et al. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut* **72**, 338–344. 10.1136/gutjnl-2022-327736 (2023).
48. Sharma, R. et al. Global, regional, and national burden of colorectal cancer and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Gastroenterol. Hepatol.* **7**, 627–647. 10.1016/S2468-1253(22)00044-9 (2022).
49. Kachuri, L. et al. Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.* **25**, 8–25. 10.1038/s41576-023-00637-2 (2024).
50. Carethers, J. M. The Jeremiah Metzger Lecture: Environmental Influences on Colorectal Cancer. *Trans. Am. Clin. Climat Assoc.* **134**, 181–199 (2024).
51. Muller, C., Ihionkhan, E., Stoffel, E. M. & Kupfer, S. S. Disparities in early-onset colorectal cancer. *Cells* **10** 10.3390/cells10051018 (2021).
52. Rosenthal, E. A. et al. Comparing ancestry standardization approaches for a transancestry colorectal cancer polygenic risk score. *Genet. Epidemiol.* **49**, e22590. 10.1002/gepi.22590 (2025).
53. Fernandez-Rozadilla, C. et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nat. Genet.* **55**, 89–99. 10.1038/s41588-022-01222-9 (2023).
54. Study, C. et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40**, 1426–1435. 10.1038/ng.262 (2008).
55. Peters, U. et al. Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology* **144**, 799–807e724. 10.1053/j.gastro.2012.12.020 (2013).
56. Hughes, E. et al. Development and validation of a breast cancer polygenic risk score on the basis of genetic ancestry composition. *JCO Precis Oncol.* **6**, e2200084. 10.1200/po.22.00084 (2022).
57. Cheng, S. et al. Validation and context-dependent effects of a prostate cancer polygenic risk score in the All of Us Research Program. *Am. J. Hum. Genet.* **113**, 392–398 (2026).
58. Mavrych, V., Alamil, M., Bolgova, O. & Dvornyk, V. Major ethnic populations are significantly differentiated at the glioblastoma multiforme candidate loci. *Int. J. Mol. Sci.* **27** 10.3390/ijms27104424 (2026).
59. Jain, P. R. et al. Can polygenic risk scores help explain disease prevalence differences around the world? A worldwide investigation. *BMC Genomic Data*. **24**, 70. 10.1186/s12863-023-01168-9 (2023).
60. Ge, T. et al. Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. *Genome Med.* **14**, 70. 10.1186/s13073-022-01074-2 (2022).
61. Arnau-Collell, C. et al. Evaluating the potential of polygenic risk score to improve colorectal cancer screening. *Cancer Epidemiol. Biomarkers Prev.* **31**, 1305–1312. 10.1158/1055-9965.epi-22-0042 (2022).
62. Vogelstein, B. et al. Cancer genome landscapes. *Sci. (New York N Y)*. **339**, 1546–1558. 10.1126/science.1235122 (2013).

Figures

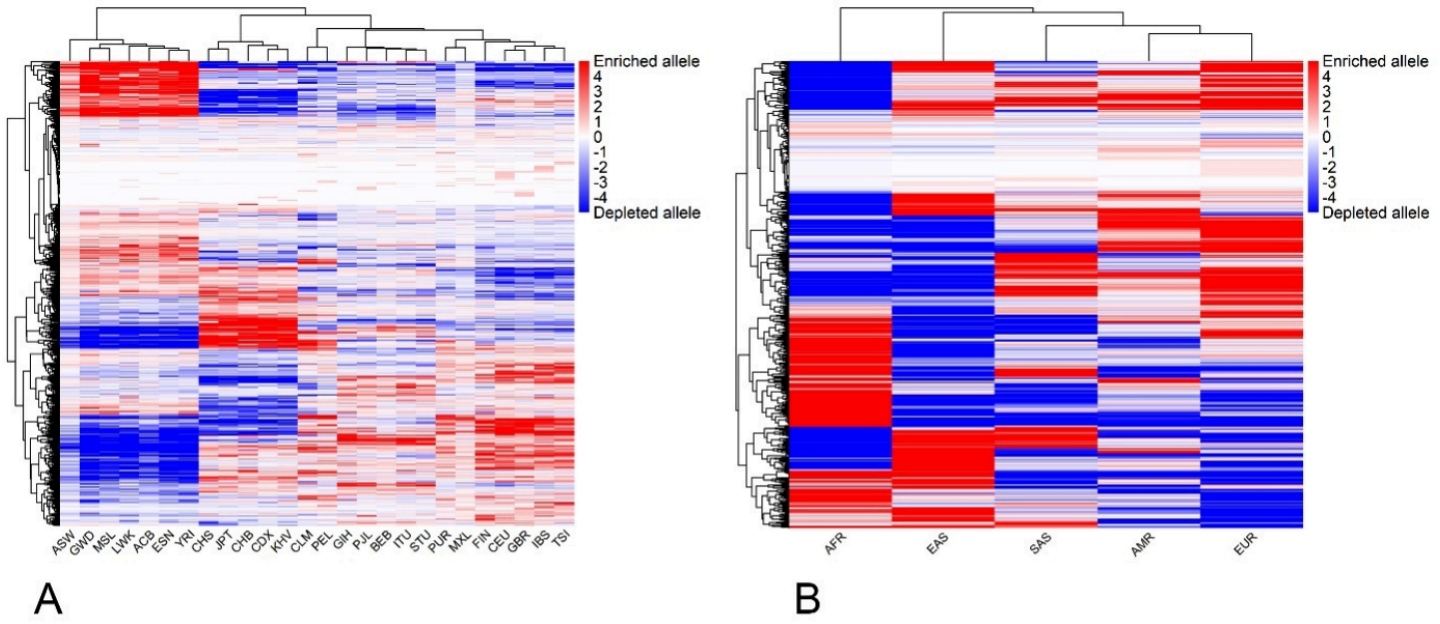


Figure 1

A heatmap of the effect allele frequencies in the studied populations. The color scale represents the depletion (blue) and enrichment (red) of the allele as compared to the average allele frequency in the whole sample (white). A – subpopulations; B – superpopulations. For the population designations, see Supplementary Table 1.

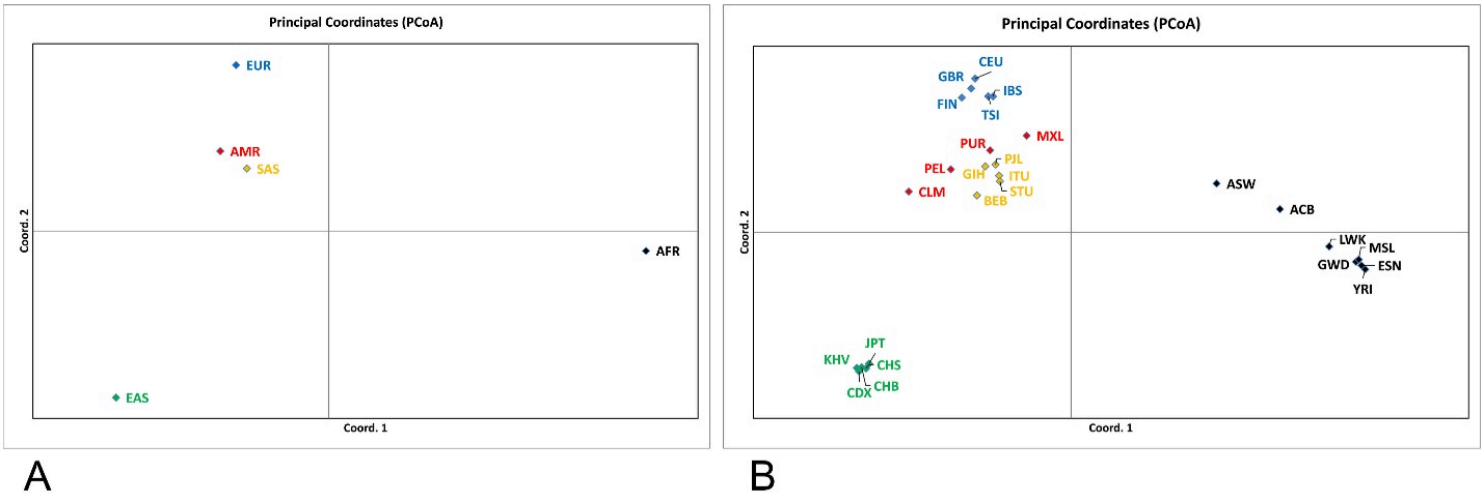


Figure 2

Principal Coordinate Analysis of the super- and subpopulations.A – superpopulations; B – subpopulations. The superpopulations and respective subpopulations are color-matched. For the population designations, see Supplementary Table 1.

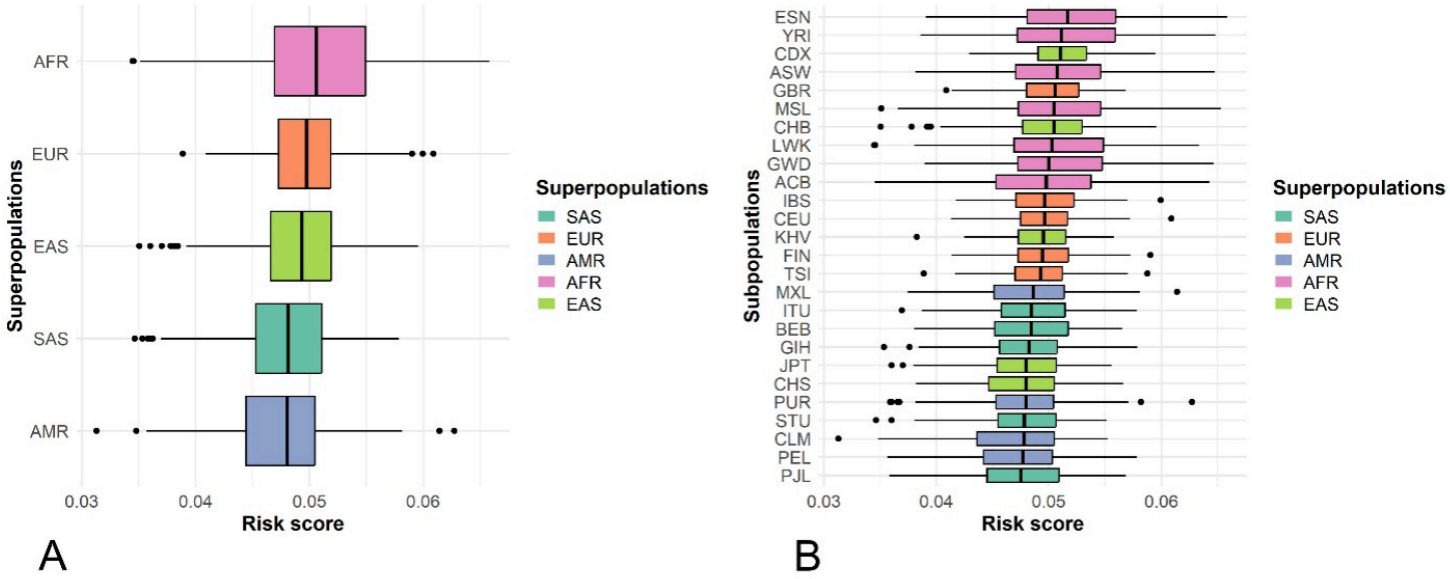


Figure 3

Box plot diagram of population-specific composite genetic risk scores for CRC. A – superpopulations; B – subpopulations. The center line of the box plot represents the median, the edges of the box indicate the 25th and 75th percentiles; the whiskers extend 1.5 times the interquartile range from these percentiles, and circles denote outliers. For the population designations, see Supplementary Table 1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarymaterial.zip