The contribution of coding variants in cancer GWAS susceptibility regions to cancer risk

Austin Hammermeister Suger

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2022

Committee:

Sara Lindstroem (Chair)

Alison Fohner

Amanda I. Phipps

Program Authorized to Offer Degree:
Public Health Genetics

University of Washington

Abstract

The contribution of coding variants in cancer GWAS susceptibility regions to cancer risk

Austin Hammermeister Suger

Chair of the Supervisory Committee:

Sara Lindstroem

Department of Epidemiology

Deepening our understanding of the genetic architecture of cancer could provide insights into cancer biology to help mitigate the health effects of increasing global cancer incidence. Elucidating the shared genetic etiology of cancer could reveal novel loci involved in cancer susceptibility. We investigated rare coding variation in genes near cancer susceptibility regions to identify genes with pleiotropic cancer associations. We used a nested case-control design to sample 195,507 participants with whole exome sequencing data from the UK Biobank cohort. We then conducted rare coding variant analyses in a generalized linear mixed-effects model framework to robustly control for population stratification and family relatedness. Our analyses identified a significant association between predicted deleterious rare coding variation in *BRCA2* and cancer diagnosis. We also found a significant rare variant association in *MC1R* and suggestive associations in several other genes. These results highlight the potential for investigating genetic contributions to cross-cancer risk using large population-based samples that include individuals representing diverse genetic ancestries.

I. Introduction

Global cancer mortality and incidence

Cancer is a leading cause of global mortality, ranking as the first or second most common cause of death before the age of 70 in 112 of 183 countries evaluated by the World Health Organization (WHO) in 2019 [1]. Global estimates from GLOBOCAN 2020 indicate approximately 19.3 million new cancer cases and close to 10 million cancer deaths in 2020 [2]. Current projections suggest a substantial increase in cancer diagnoses over the next two decades, with the increased cancer incidence affecting all countries and regions [2,3]. This is expected to be accompanied by increased cancer-related morbidity and mortality, which could have a disproportionate burden on countries with limited healthcare resources [2,4]. Expanding our understanding of cancer biology and etiology is an essential part of improving cancer prevention, screening, and treatment strategies to attenuate the effects of increasing cancer incidence.

Genetic susceptibility to cancer and pleiotropy

Evidence supporting the foundational role of genetic variation in cancer susceptibility has continuously expanded over the past five decades. Early studies identified mutations in the *Rb* gene that predispose individuals to the development of retinoblastoma and thereby elucidated the impact of germline mutations on cancer risk [5,6]. Subsequent studies uncovered numerous genomic loci (e.g., *TP53*, *APC*, *PTEN*, *ATM*, *CHEK2*, *BRCA1*, *BRCA2*, and *PALB2)* later shown to have substantial influences on risk for a variety of cancer types [7–14]. A more recent

study of 200,000 individual twins from population-based registers in Nordic countries estimated an overall cancer heritability of 33% (95% CI, 30%-37%) [15]. The heritability estimates varied substantially by cancer type, from 0.15 for colon cancer to 0.57 for prostate cancer. These data support an integral role of genetic factors in the etiology of cancer and the importance of elucidating the genetic architecture of cancer to improve our understanding of that etiology. The genetic etiology of cancer is complex, with genetic variation in known high-impact loci accounting for only a small proportion of the total heritability. The remaining heritability of cancer is thought to be contributed by numerous loci with weaker individual effects [16]. Over the past two decades, hundreds of genome-wide association studies (GWAS) have been conducted to identify genetic loci associated with cancer risk. As of August 2021, there were over 2,400 unique single nucleotide polymorphisms (SNPs) with a genome-wide significant association ($p < 5 \times 10^{-8}$) with one or more cancer types in the GWAS Catalog [17]. These associations represent at least 28 distinct cancer types, including common cancers such as breast, prostate, and colorectal, as well as rarer cancers such as cancers of the central nervous system and esophageal cancer. Of the cancer-associated SNPs identified by GWAS, approximately 190 have reported associations with multiple cancer types, suggesting loci with pleiotropic effects across cancer types.

Cancer pleiotropic loci are supported by studies demonstrating modest genetic correlations between common cancer types [18–21]. Meta-analyses of GWAS have leveraged pleiotropic effects to identify loci with associations shared by multiple cancer types, including breast, ovarian, prostate, and endometrial cancers [22,23]. Focusing on pleiotropic effects may allow the detection of cancer risk loci that have not been detected by previous genetic studies

of a single cancer type. Deepening our understanding of loci influencing cross-cancer susceptibility will help extend current knowledge of shared biological mechanisms of carcinogenesis. This knowledge could enhance cancer screening, identify new drug targets, and advance precision medicine therapeutic approaches. Knowledge of cross-cancer susceptibility loci may be particularly relevant for improvements in screening and treating individuals at higher risk of developing multiple primary tumors [24,25]. Estimates of the global prevalence of multiple primary cancers are between 2 and 17%, with differences in reporting practices resulting in a wide range of estimates. Only a small proportion of these cases of multiple primary cancers can be attributed to hereditary cancer syndromes, such as hereditary breast and ovarian cancer, Li-Fraumeni, and Lynch syndrome. Enriching our understanding of cancer risk loci with pleiotropic effects will better position us to screen and treat multiple primary cancers.

Rare genetic variation and large population-based sequencing data sets

Many of the previous genome-wide efforts to identify cancer risk loci have focused on investigating associations between individual cancers and individual common genetic variants. Although these variants are common in populations, they make up a small proportion of total human genetic variation. A large majority of genetic variants observed in human populations are individually rare, with the 1000 Genomes Project reporting that out of the ~88 million variants they identified in their 2,504 ancestrally diverse samples, ~64 million of them had alternative allele frequencies < 0.5%, while only ~8 million had alternative allele frequencies > 5% [26]. Rare variants make up an even larger proportion of variants in coding regions, with a

recent analysis in the whole-exome sequencing (WES) of 454,787 individuals from the UK Biobank (UKBB) cohort reporting that, of the 12.3 million variants they identified, 99.6% had minor allele frequencies (MAF) < 1% [27]. Studying these rare variants provides an opportunity to discover novel loci with associations pleiotropic to cancer as well as to identify the causal variants in known susceptibility loci.

Investigating rare coding variation is especially appealing because many popular rare variant analysis methods involve grouping variants into sets to improve statistical power, and defined coding units like genes are intuitive definitions for variant sets [28]. Exploring associations with rare coding variants is also attractive because certain portions of coding regions are critical to gene function and have a low tolerance for mutation [29,30]. Rare variants in these coding regions may be more likely to have functional consequences and affect disease than variants in non-coding regions. Numerous studies have identified rare germline coding variants associated with risk for a variety of individual cancer types, including breast, lung, pancreas, and ovary [31–35]. Many of these rare coding variants were found in genes near cancer susceptibility loci identified by prior GWAS cancer risk. This observation is supported by the findings of a recent broad gene–trait association analysis using WES data from 454,787 UKBB participants where rare variant trait associations had 3.8 to 11.4 fold enrichment within 1 Mb of a GWAS locus for that same trait [27]. This enrichment of rare variants in GWAS loci presents the possibility of leveraging the results of previous cancer GWAS to narrow down regions of interest in the search for rare coding variants that influence cancer risk.

The generation of sequencing datasets for large population-based cohorts, such as UKBB, provides the possibility to investigate the impacts of previously understudied rare

genetic variation on human disease. These large population-based datasets also provide rich phenotypic characterization and harmonization with linkage to genetic data [36,37]. The availability of this genetic and phenotypic data to academic research groups allows for the conduct of genetic studies at a scale that would be otherwise infeasible. For example, rare variant analyses for variants with modest to intermediate effects are only feasible in large studies because the rarity of the variants in the population. The accessibility of these biobank resources also facilitates the study of loci with pleiotropic effects across cancer types by allowing the study of large numbers of individuals with cancer diagnoses at a diverse set of tissue sites, in addition to individuals without cancer diagnoses.

Diversity in genetic studies, complex population structure, and mixed models

One of the most concerning issues in modern human genetic studies is the limited diversity in ancestry among studied individuals. An overwhelming majority of the total number of participants in GWAS are of European ancestry, with the GWAS Diversity Monitor reporting that only 4.18% of total GWAS participants in discovery samples are of non-European or mixed ancestry as of May 2022 [38]. This lack of diversity obstructs our ability to understand the architecture of human disease, minimizes the efficacy of genetic risk prediction models in non-European populations, and exacerbates existing global heath disparities [39–41]. When individuals with non-European ancestry are recruited into a study, they are often excluded from the main analyses to simplify analysis and minimize the confounding effects of population stratification. While controlling for population stratification is a major concern in GWAS, there are scientific

and ethical imperatives to increase diversity as well as new tools in development to help scientists conduct robust analyses in ancestrally diverse samples [39].

Controlling for population stratification (and cryptic relatedness) may be of even greater importance for rare variant analyses since rare variants tend to be more strongly influenced by population structure than common variants [42]. The most common method of accounting for population structure in GWAS, principal-component analysis (PCA), works well for low-dimensional population structure but not for the high-dimensional population and family structure that influence rare variants [28]. Furthermore, research has suggested that high-dimensional population structure exists within large biobanks, such as the UKBB cohort, and could have implications for inferences made from genetic studies adjusting for population stratification solely using PCA [43]. Another method to control for population structure is generalized linear mixed-effects models (GLMM), which can adjust for population structure, cryptic relatedness, and family structure simultaneously by adjusting for empirical kinship estimates using a genetic relationship matrix (GRM) [44]. In the past, GLMMs have been unpopular due to high computational costs and methodological challenges using GLMMs to assess genetic associations in case-control designs [45]. However, recent developments now allow GLMMs to be integrated with popular rare variant analysis methods to test binary traits in large imbalanced case-control studies[46–50]. This makes it feasible to conduct large rare variant analyses while also including ancestrally diverse samples and robustly adjusting for population stratification.

Goals of the project

This project aims to study the contribution of coding variation within known cancer susceptibility regions in cancer risk using gene-based variant set analysis in a GLMM framework. This strategy will provide an opportunity to identify novel candidate cross-cancer susceptibility genes not pinpointed in past studies of common genetic variation. Focusing the analysis on regions near known GWAS cancer susceptibility regions increases the chances of identifying loci where rare coding variants contribute to pleiotropic risk across cancer types. Results from gene-based analyses will be used to investigate the potential biological process involvements of cancer associated genes.

II. Methods

UK Biobank data

The UKBB is a large population-based prospective cohort study of 500,000 individuals in the United Kingdom. The cohort participants were recruited from 2006 to 2010 at 22 assessment centers across the UK, and their ages ranged from 40 to 69 years at recruitment [51]. Upon recruitment, participants completed questionnaires, were interviewed, and underwent a physical examination. Blood samples were collected from all participants to be used for genomic and biochemical assays. Genome-wide genotype data has been available for all participants in the UKBB cohort since July 2017. The first 50,000 participants were genotyped using the Affymetrix UK BiLEVE Axiom array, and the remaining 450,000 participants were genotyped using the Affymetrix UK Biobank Axiom Array [52]. WES data for 50,000 UKBB participants was released in March 2019, and WES data on an additional 150,000 participants

was released in October 2020 [53]. Exome sequences were captured using the IDT xGen Exome Research Panel v1.0 plus supplemental probes targeting 39 Mbp of the human genome, corresponding to 19,396 genes [54]. The first 50,000 participants were sequenced with a different IDT v1.0 oligo lot than the following 150,000 samples. The first 50,000 sequenced participants were selected based on the availability of MRI data and extended baseline measurements, while the subsequent 150,000 samples were randomly selected from the full cohort. The age, sex, genetic ancestry, and cancer diagnosis distributions for the first 50,000 individuals suggest that this group is representative of the full cohort [54]. Participant health records, including cancer registry records, are periodically linked with UKBB data. Cancer diagnosis records for UKBB participants are provided to the UKBB by the Medical Research Information Service NHS for individuals living in England and Wales and by the Information Services Division NHS Scotland for individuals residing in Scotland [55]. The latest cancer registry record linkage with UKBB data was June 25, 2021. UK Biobank participant data is deidentified prior to being made accessible to researchers. This research was conducted using the UK Biobank Resource under Application Number 70925.

Sample selection

The initial study population was the full UKBB cohort of 502,415 participants. Five individuals were immediately excluded from the eligible participants due to recent withdrawal from the UKBB cohort. A nested case-control design was then used to select eligible cancer cases and controls within the remaining 502,410 UKBB cohort participants. Cancer registry records were used to identify subjects with at least one prevalent or incident malignant cancer

diagnosis and determine their primary cancer diagnosis. Here, a primary cancer diagnosis was defined as a participant's earliest diagnosed invasive or *in situ* cancer based on the diagnosis date reported by a cancer registry. If a participant had an invasive and an *in situ* diagnosis record, the earliest invasive cancer diagnosis record was considered the primary cancer. Eligible cases were individuals with an ICD9/10 code for an invasive or *in situ* primary cancer diagnosis at a specified or well-defined site. A total of 106,144 eligible cases were identified in the full UKBB cohort based on these criteria. Eligible controls were individuals without a cancer registry diagnosis record who also self-reported no previous cancer diagnoses. A total of 384,337 eligible controls were identified in the full UKBB cohort based on these criteria.

Only UKBB cohort participants with 200k release WES data were included in the analysis, reducing the number of potential cases to 41,709 and the number of potential controls to 154,046. Of these 195,755 individuals, 166 individuals were excluded due to missing genotype data. An additional 59 individuals were excluded due to discordant self-reported and genetically determined sex, as this discrepancy could indicate potential genotype data quality issues. Pairwise kinship estimates from genotype data identified 23 pairs of individuals among the remaining 195,530 subjects as monozygotic (MZ) twins. One individual from each pair of twins was excluded. If the MZ twin pair were both controls, the excluded individual was chosen at random. If the MZ twin pair included a case and a control, the control was excluded. If the MZ twin pair included two cases, the case with the more common primary cancer diagnosis site was excluded. An additional exclusion criterion of >10% missingness rate across variant sites in the WES data was applied. However, no individuals were excluded under this criterion.

Cancer GWAS susceptibility regions

Genomic regions known to influence cancer susceptibility were identified using SNPs with reported associations to any cancer in the GWAS Catalog [17]. A list of 9,988 SNPs with associations to the trait "cancer" was downloaded from the GWAS Catalog. SNPs with a reported cancer associations not reaching genome-wide significance ($p < 5 \times 10^{-8}$) were excluded from this list. SNPs were also excluded if their reported association was with a non-diagnosis cancer trait (e.g., response to radiation, cancer survival time). Duplicate SNP-trait association for each cancer type were removed so that each SNP was included only once within the list for each cancer type. SNPs with reported associations to multiple cancer types or an association explicitly reported as a cancer pleiotropy were labeled as pleiotropic and the remaining duplicates were removed such that every SNP in the list was unique. After these filtering steps, 2,491 unique cancer-associated SNPs remained (Figure 1), and 194 (7.8%) were associated with multiple cancer types (Figure 2). These 2,491 cancer-associated SNPs represent at least 28 unique cancer types (Table 1). Genomic positions for these SNPs were used to generate 534 non-overlapping genomic regions that contain all cancer-associated SNPs (Figure 3). These cancer GWAS susceptibility regions were created by placing each cancer-associated SNP at the midpoint of a 1 Mb window and then combining overlapping windows into a new larger window. All figures illustrating the genomic locations of cancer GWAS SNPs and susceptibility regions were created using the PhenoGram visualization tool [56].

WES capture target gene annotations

The UKBB WES with the IDT xGen Exome Research panel v1.0 targeted 204,829 capture targets across 39 Mbps of the human genome. The Matched Annotation from NCBI and EMBL-EBI (MANE) v1.0 annotation database was used to establish gene annotations for these capture targets [57]. MANE exon and gene annotations were used in combination to annotate the UKBB WES capture targets. Discrepancies between WES capture target gene annotations generated by the two methods (exon-based and gene-based) were resolved by manual curation. Using MANE, a total of 199,859 (192,636 autosomal) WES capture targets, corresponding to exons from 18,662 (17,848 autosomal) genes, were annotated.

Variant annotation and filtering

Variants present in the UKBB WES 200k release data were annotated for functional impacts on nearby genes using SnpEff [58]. SnpSift was used to identify variants with SnpEff annotated "HIGH" or "MODERATE" impacts, and these variants were included in gene-based variant sets [59]. Examples of variants with moderate SnpEff annotated effects include in-frame insertions and deletions, missense variants, and splice region variants. Examples of variants with high SnpEff annotated effects include stop gain and loss variants, start loss variants, splice donor and acceptor variants, rare amino acid variants, and frameshift variants. Variants with variant missingness < 10%, minimum read depth coverage ≥ 7, Hardy-Weinberg equilibrium p-value > $1x10^{-15}$, QUAL score ≥ 30, and variants with at least one heterozygous sample with an allele balance threshold > 0.15 were included in variant sets. These filters were applied using PLINK 2.0 and bcftools [60,61]. Monomorphic and multiallelic variants were also excluded from variant sets.

Principal components and genetic relatedness matrix marker data set

The UK Biobank computed the first 40 principal components (PCs) from 147,603 high quality LD pruned genetic markers across 488,175 UKBB participants using fastPCA and made them accessible to researchers [36,62]. The first 20 PCs were used as covariates in models to account for low dimension population structure (Supplementary Figures 1-11). A data set of 146,521 high-quality LD pruned variants with MAF ≥ 0.01 and 106,922 high-quality variants with MAF < 0.01 from UKBB genotype data was created for GRM construction and rare variant variance ratio estimation, respectively.

Statistical analyses

Set-based rare variant association analyses for gene-based variant sets were performed using SAIGE-GENE v0.45 [49,50]. Let *N* represent the number of individuals in the sample, and let *q* represent the number of variants tested in each variant set. SAIGE-GENE uses a generalized linear mixed model (GLMM), which can be expressed as:

$$g(\mu_i) = X_i\alpha + G_i\beta + b_i$$

Where the link function *g* is the identify function when testing for continuous traits with an error term e $\sim N(0, I\sigma_e^2)$ or a logistic function when testing for binary traits. μ_i is the mean of the trait being tested. $X_i$is a vector of covariates including the intercept. G_i is the minor allele dosage for a given variant (0, 1, 2). α is a vector of fixed effects covariates and β is a *q* x 1 vector of genetic effects. b_i is a random effect distributed as $N(0, \tau\psi)$. τ is the additive genetic variance and ψ represents the *N x N* GRM. SAIGE-GENE also accounts for unbalanced case-

control ratios when testing binary traits by using saddlepoint approximation and efficient resampling [63,64].

A sparse GRM using 2,000 randomly sampled markers from the GRM data set was constructed. This sparse GRM was created such that GRM elements below a relatedness coefficient of 0.125 were zeroed out. GRM elements for pairs of individuals with up to third-degree relatedness were preserved in the sparse GRM. The null GLMM was fit using the PQL method and average information restricted maximum likelihood (AI-REML) algorithm to iteratively estimate ($\hat{\alpha}$, $\hat{b}$, $\hat{\tau}$) under a null hypothesis of β = 0 [49,50,65–67]. Cancer case status was defined as a binary variable in the model, and the fixed effects covariates included in the null model were age at recruitment, genetically determined sex, WES batch, and the first 20 PCs. When fitting the null model, the sparse GRM was used to estimate single-variant score statistic variances for extremely rare variants using the ratio of the variance of randomly selected rare markers in the sparse GRM and the full GRM. Variance ratios were estimated separately for variants in minor allele count (MAC) categories of 1, 2, 3, 4, 5, 6-10, 11-20, and >20. The leave-one-chromosome-out (LOCO) method was used to avoid the proximal contamination problem, and maximize statistical power, when using genetic markers to adjust for sample relatedness [45]. GRMs were constructed separately for each chromosome by leaving markers on the chromosome to be tested out of the chromosome specific GRM.

After fitting the null model, gene-based variant sets were established from the variants that had passed filtering criteria. All gene-based tests were conducted using SKAT-O, which uses a linear combination of burden (Q_{burden}) and SKAT statistics (Q_{SKAT}) to provide robust power [46–48]. The SKAT-O test statistic (Q_{SKAT-O}) can be written as:

$$Q_{SKAT-O} = (1-\rho)Q_{SKAT} + \rho Q_{burden}\ with\ 0 \leq \rho \leq 1$$

When performing SKAT-O tests, SAIGE-GENE uses a set of eight values for ρ (0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1) to estimate the minimum *p*-value using numerical integration.

A set of potential cancer susceptibility genes were defined as 7,421 autosomal genes that overlapped with the previously defined 525 cancer GWAS susceptibility regions on autosomes. Those 7,421 genes corresponded to 79,657 WES capture target regions within the UKBB data set. For each gene, two sets of variants were created for testing. One set included coding variants that had predicted high and moderate impacts based on SnpEff annotations. The second set included coding variants that had high impacts based on SnpEff annotations.

Missing dosages for variants were mean imputed in the analysis. Variants were weighted by MAF using a *Beta*(1, 25) distribution. Any variants within a gene-based set with MAC ≤ 10, here termed ultra-rare variants, were collapsed into a presence or absence dosage variable. Individuals with an alternate allele dosage ≥ 1.5 for any of the ultra-rare variants in the set were assigned a dosage of 2 in the presence or absence dosage variable. Individuals with an alternate allele dosage ≥ 0.5 and < 1.5 for any of the ultra-rare variants in the set were assigned a dosage of 1 in this presence or absence dosage variable. All other individuals received a dosage of 0 for the presence or absence dosage variable. Any gene with fewer than two variants in its set was excluded from the analysis. For the moderate and high impact variant sets, 6,597 SKAT-O tests were performed. For the high impact variant sets, 6,046 SKAT-O tests were performed. The total number of SKAT-O tests conducted was 12,643, yielding a Bonferroni corrected significance threshold of 3.95×10^{-6} ($\alpha = 0.05/12{,}643 = 3.95 \times 10^{-6}$). Single variant tests on all variants included in the sets were conducted concurrently with the SKAT-O

tests for high and moderate impact variant sets. Ultra-rare variants were not included in the single variant tests. A total of 148,938 single variant tests were conducted, resulting in a Bonferroni corrected significance threshold of 3.35×10^{-7} ($\alpha = 0.05/148{,}938 = 3.35 \times 10^{-7}$). Conditional tests were performed to interrogate the influence of significant single variants on their gene set association and nearby single variant signals. Linkage disequilibrium (LD) was also estimated for significant and suggestive single variant associations in the same variant sets.

A gene-set enrichment analysis (GSEA) for biological process enrichment among significant and suggestive genes with $p < 1 \times 10^{-3}$ was conducted using the PANTHER Overrepresentation Test with the PANTHER version 16 classification system [68–70]. The GSEA was performed using Fisher's exact test in the GO biological process complete data set. All *p*-values were corrected for multiple testing using the Bonferroni correction method. The reference list for the overrepresentation test included only the genes tested in the gene-set analysis.

IRB Human Subjects Research Determination

The University of Washington Human Subjects Division reviewed the proposed project protocol (IRB ID: STUDY00013674) and determined that the proposed activities do not involve human subjects.

III. Results

Study sample

There were 41,670 cancer cases and 153,837 controls for the main analyses after applying exclusion criteria. The demographic characteristics of these 195,507 subjects are

displayed in Table 2. On average, cases were older, with a mean age at recruitment of 59.7 years (SD = 7.2 years) compared to controls with a mean age at recruitment of 55.5 years (SD = 8.1 years). The distribution of primary cancer sites among the 41,670 cancer cases is presented in Table 3. After grouping cancers by ICD9/10 codes, we identified at least 48 distinct cancer diagnosis sites among the cases and 14 *in situ* diagnosis sites. The five most common cancers among cases in the study sample, including invasive and *in situ* diagnoses, were non-melanoma skin cancers (27.64%), breast cancer (18.51%), prostate cancer (11.67%), colorectal cancer (6.38%), and cervical cancer (4.67%). Together, these cancer types accounted for 66.87% of the 41,670 cases in the study sample. When including only invasive diagnoses, the five most common cancers among cases in the study sample were non-melanoma skin cancers (26.98%), breast cancer (16.45%), prostate cancer (11.67%), colorectal cancer (6.38%), and melanoma (4.19%). These five invasive cancer diagnoses are consistent with estimates of the most prevalent cancers in the general UK population.

SKAT-O tests in cancer GWAS susceptibility region genes

After conducting 6,597 gene-based high and moderate impact variant SKAT-O tests, no genes were found to be significantly associated with cancer diagnosis (Figure 4). Results for all genes with $p < 1 \times 10^{-3}$ in these SKAT-O tests are presented in Table 4. The two genes with the strongest suggestive associations with cancer diagnosis were Mucin Like 3 (*MUCL3*; $p = 8.27 \times 10^{-6}$) and melanocortin-1 receptor (*MC1R*; $p = 9.5 \times 10^{-6}$). The quantile-quantile (Q-Q) plot for these SKAT-O tests shows overall deflation of the test statistics ($\lambda=0.898$) but inflated test statistics for the genes with stronger associations (Figure 5). After conducting 6,046 gene-based

high impact variant SKAT-O tests, *BRCA2* was found to be significantly associated with cancer diagnosis ($p = 1.09 \times 10^{-6}$), with no other gene-based sets showing significant association (Figure 6). Results for all genes with $p < 1 \times 10^{-3}$ in these SKAT-O tests are presented in Table 5. A suggestive association with cancer diagnosis was found for *BRCA1* ($p = 6.08 \times 10^{-6}$) in the high impact variant set tests. The Q-Q plot for these SKAT-O tests shows adequate control for genomic inflation ($\lambda = 0.998$) but overall deflation of test statistics (Figure 7).

Single variant tests in cancer GWAS susceptibility region genes

After performing 149,628 single variant tests of high and moderate impact variants, a variant on chromosome 16 at position 89,919,709 was found to be significantly associated with cancer diagnosis ($p = 1.83 \times 10^{-18}$), with no other significant associations detected (Figure 8). Results for all variants with $p < 1 \times 10^{-5}$ in single variant tests are presented in Table 6. The significant variant was a C/T (major allele/minor allele) missense variant in *MC1R* with the designation rs1805007. rs1805007 had an overall MAF of 0.096 in the study sample and a MAF of 0.107 among cases, compared to a MAF of 0.093 among controls. The strongest suggestive single variant association was for a variant on chromosome 16 position 89,919,736 ($p = 6.5 \times 10^{-7}$). This suggestive variant is a C/T missense variant in *MC1R* with the designation rs1805008. rs1805008 had an overall MAF of 0.083 in the study sample, with a MAF of 0.094 among cases and a MAF of 0.081 among controls. No individuals in the study sample had a TT haplotype for rs1805007 and rs1805008 (D' = 1), but alleles for the two variants do not show a high degree of correlation ($r^2 = 0.0097$). The Q-Q plot for the single variant tests shows adequate control for genomic inflation ($\lambda = 1.002$) and no systematic inflation of test statistics (Figure 9).

Conditional analyses

When conditioning the high and moderate impact *MC1R* variant set test on the significant and suggestive single variant associations rs1805007 and rs1805008, the suggestive association with cancer diagnosis was nullified ($p = 0.224$). Conditioning this *MC1R* variant set test on either rs1805007 ($p = 0.00264$) or rs1805008 ($p = 0.00054$) also removed the suggestive association. When conditioning the single variant test for rs1805007 on rs1805008, the association with cancer diagnosis remained highly significant ($p = 8.63 \times 10^{-21}$). Conditioning the single variant test for rs1805008 on rs1805007 also yielded a significant association ($p = 2.60 \times 10^{-9}$). Detailed results from the conditional analyses are presented in Table 7.

Gene set enrichment analyses

The significant and suggestive genes (SKAT-O $p < 1 \times 10^{-3}$) identified in high or high and moderate impact variant set tests were not significantly overrepresented in GO biological processes after Bonferroni correction.

IV. Discussion

Cancer risk associated genes

BRCA2 was found to be significantly associated with cancer diagnosis when including predicted high impact variants in the SKAT-O tests. *BRCA1* showed the strongest suggestive association in these variant set tests. *BRCA2* and *BRCA1* are tumor suppressor genes involved in

the homology-directed repair of double-stranded DNA breaks [71–73]. Germline mutations in the *BRCA* genes are known to have substantial impacts on an individual's lifetime risk of breast and ovarian cancer [74–76]. Mutations in these genes have also been associated with an increased risk of prostate, pancreatic, and endometrial cancers [77–81]. Collectively, these cancers accounted for 37.79% of the 41,670 cases in the study sample. In addition, germline mutations in *BRCA2* have strong associations with certain forms of melanoma as well as putative associations with lung, esophageal, cervical, stomach, and non-melanoma skin cancers [82–87]. These five cancers accounted for an additional 42.63% of the 41,670 cases in the study sample. The cancer GWAS susceptibility region overlapping *BRCA2* included 10 SNPs associated with breast, colorectal, lung, non-melanoma skin cancers, and cancer pleiotropy. The cancer GWAS susceptibility region overlapping *BRCA1* included 20 SNPs associated with breast cancer, esophageal cancer, ovarian cancer, and cancer pleiotropy. The observed significant association between rare coding genetic variation in *BRCA2* and cancer diagnosis is consistent with its established impact on cancer risk. These results also support evidence that *BRCA2* is involved in the genetic etiologies of a broader range of cancers than *BRCA1*.

MUCL3 had the strongest suggestive association in the high and moderate impact variant set tests. Associations between *MUCL3* germline mutations and cancer have not been well studied. There have been investigations into somatic mutations and differential gene expression of *MUCL3* in tumors. A recent study of lung adenocarcinoma identified overexpression of *MUCL3* in tumor cells with invasive mucinous adenocarcinoma morphologies when compared to tumor cells with non-terminal respiratory unit type lung adenocarcinomas [88].

MC1R had the second strongest suggestive association in the high and moderate impact variant set tests and the sole significant variant association in single variant tests. *MC1R* is involved in melanin pigmentation of human skin through its interaction with the melanocyte-stimulating hormone and may have a role in stimulating DNA repair [89,90]. Studies have established that germline variants in *MC1R* are associated with risk of skin cancers [91–94]. The single variant analyses identified that rs1805007 was significantly associated with cancer diagnosis in the study sample. The T allele for rs1805007 has a scaled CADD score of 25.2, suggesting that the predicted deleteriousness is in the top 1% of deleteriousness for all human genetic variants [95]. The R151C *MC1R* exon 3 variant rs1805007 has eight reported genome-wide significant ($p < 10^{-8}$) associations with non-melanoma skin cancers and two with melanoma in the GWAS Catalog [17]. The single variant analyses also identified that rs1805008 has a suggestive association with cancer diagnosis in the study sample. The T allele for rs1805008 has a scaled CADD score of 22.5, suggesting that the predicted deleteriousness is in the top 1% of deleteriousness for all human genetic variants [95]. The R160W *MC1R* exon 3 variant rs1805008 has one reported genome-wide significant association with melanoma in the GWAS catalog. The nullification of the association in the conditional test for the high and moderate impact *MC1R* variant set suggests that rs1805007 and rs1805008 are driving the observed suggestive association. Conditional single variant analyses suggest that these variants have independent associations with cancer diagnosis. This is supported by the low correlation between the two loci that was observed in the study sample ($r^2 = 0.0097$). The LD observed between these two variants among study participants included in the analysis is concordant with the LD observed in samples from England and Scotland in the 1000 Genomes Project (r^2 =

0.0084; D' = 1) [26]. There is little existing evidence that germline genetic variation in *MC1R* is associated with risk for non-skin cancers. Skin cancer cases made up 32.9% (13,725) of the 41,670 cases in the study sample, which could explain the identification of *MC1R* variant associations with cancer diagnosis. Furthermore, the relatively high MAFs of rs1805007 and rs1805008 meant that single variant tests had far greater power to detect associations from these variants compared to other variants tested.

Some inflation of test statistics was observed for genes with the strongest associations to cancer diagnosis in the high and moderate impact variant set SKAT-O tests (Figure 5). The other gene-based variant set tests, and single variant tests appeared to have adequate control for genomic inflation (Figures 7 & 9). The inflation seen in the high and moderate impact set tests could indicate issues in controlling type I error for that set of SAIGE-GENE SKAT-O tests. The authors of SAIGE-GENE indicated that they observed inflated type I error rates for certain phenotypes and using a greater number of markers in GRM construction could improve model performance [49,50]. The GRM incorporated to account for population structure in our analyses was created using 147,603 LD pruned common variant markers. Including additional markers in the GRM might better control for the complex population structure in the UKBB cohort (Supplementary Figures 1-11). It is also important to note that the selection of genes and variants for testing was not agnostic. Genes were chosen based on their proximity to known cancer GWAS susceptibility loci, and variants were selected based on predicted functional impact. This could explain some of the observed early deviation of test statistics from the null distribution (Figure 5).

Genetic correlations across cancer types

Prior research has revealed modest genetic correlations across a variety of cancer types, with the highest estimated shared heritability for small groups of cancers [18–21]. Evidence suggests moderate genetic correlations between certain cancer types, including lung and head and neck cancers, breast and ovarian cancers, and colorectal and lung cancers [18]. Weaker genetic correlations have also been observed between other pairs of cancer, such as prostate and head and neck, prostate and thyroid, and lung and breast cancers [18,20]. However, significant genetic correlations between most pairs of cancer types have not yet been identified. These observations support the existence of some loci with pleiotropic effects across certain groups of cancers but may not support the existence of a substantial number of loci with pleiotropic effects across all cancer types.

A major limitation of prior cross-cancer genetic correlation studies and the present analysis is that certain cancer types are far more common in the population than others. This presents challenges for studies estimating the shared heritability between certain cancer types as well as for studies investigating genetic variation associated with these cancers. Analyses of cross-cancer genetic correlations either do not have sufficient sample sizes to study rare cancer types or have lower power to detect correlations among those rarer cancer types. In our analyses, the genes tested were defined by regions near cancer-associated SNPs from the GWAS Catalog, where the three most common cancers (breast, prostate, and colorectal) contributed 49.5% of the SNPs (Table 1) [17]. Furthermore, the five most common invasive cancer diagnoses in the study population represented 65.67% of all cases included in the analysis. As a result, associations between coding variants and these common cancer types could be

obscuring the associations between coding variants and rarer cancer types. More robust assessments of the shared genetic architecture of rare cancer types will likely require larger population-based cohorts that include a sufficient number of rare cancer cases.

Representativeness of the UK Biobank cohort

The overall recruitment response rate among individuals invited to participate in the UKBB was very low, with only around 5.5% of invited individuals participating. The potential "healthy volunteer" selection bias in the UKBB cohort has raised concerns and sparked debate among epidemiologists with numerous commentaries and articles published on the topic [96–98]. Comparisons between the UKBB cohort and the general UK population have suggested the cohort is not fully representative of the sampled population [99]. UKBB participants are generally healthier and less likely to live in socioeconomically disadvantaged areas. Participants also have substantially lower all-cause mortality and approximately 10-20% lower cancer incidence rates than the general UK population. There have been investigations into if and how this lack of representativeness could affect etiological inferences made from the cohort [100,101]. These studies report mixed conclusions, and more work investigating the influence of non-representativeness in the UKBB cohort is needed.

An additional concern is the lack of ancestral diversity in the UKBB cohort relative to the UK population. Although individuals from diverse self-reported backgrounds were included in our analyses, 183,389 (93.80%) of the 195,755 individuals in the study sample self-reported their background as White. For comparison, around 95% of the UK population self-reported as White in the 2001 Census and approximately 91% self-reported as White in 2011 Census [99].

Increasing the ancestral diversity of genetic studies will help us better understand the shared genetics of cancer. It will also help us make epidemiologic discoveries more applicable to the populations of low and middle-income countries that are facing the most severe consequences of increasing cancer incidence [2,4]. This emphasizes the need for large cohorts that are more representative of diverse human genetic ancestries.

Limitations of rare variant analyses

There are many limitations to rare variant analyses using exome data. One limitation is that single variant tests often do not have adequate statistical power to identify associations between traits and rare variants. In our study, the strongest single variant associations detected were more common rare variants with MAF close to 1%. Variant set tests such as SKAT-O are commonly used to account for low power with single variant testing [47,48]. However, statistical power for SKAT-O tests is closely related to the proportion of variants with causal effects on the trait included in the variant set. If too few causal variants are included in the variant set, the test will suffer from low power. Therefore, power can be increased by including the variants most likely to have a causal effect on the trait using predicted variant impact filtering. Yet, applying filters that are too stringent can result in causal variants being excluded from variant sets. As a result, meticulous selection of variant filtering criteria is essential to maximizing test power. Our analyses used predicted impact annotations from SnpEff which categorizes predicted variant impacts into four levels, “HIGH”, “MODERATE”,” LOW”, and “MODIFIER” [58]. Using additional variant effect prediction data sets in combination with SnpEff could provide more granular variant function information that could be used to improve statistical power.

More comprehensive variant effect information would also allow for direct weighting of variants based on predicted deleteriousness. This could improve power relative to methods such as indirectly weighting variants based on deleteriousness using MAF.

A second set of limitations for rare variant analyses using WES data is that the coding regions of most genes are mutationally constrained by strong purifying selection [29,30]. This means that very large sample sizes are required to observe enough high-impact variants to perform set-based testing across all genes. In our study sample of 195,507 individuals, 824 and 1,375 genes were excluded from the high and moderate and high impact variant set tests, respectively, due to low variant counts. Structural and copy number variants are also known to be involved in cancer susceptibility, with known structural variation in *BRCA* and other cancer predisposition genes [102,103]. Variants in non-coding regions could also impact gene regulation influence cancer risk. These types of genetic variation were not captured in our analyses. Whole genome sequencing (WGS) data could ameliorate these issues as structural variation and variants in non-coding regions of genes could be measured. Due to technical limitations of the modeling software, our analyses also excluded multiallelic variants, which are estimated to comprise close to 19% of all coding variants in the UKBB 200k WES dataset [37]. Incorporating additional variant types into variant set tests would increase the number of genes that could be assessed and more comprehensively examine potential relationships between genetic variation and cancer risk.

Future directions for studying cancer pleiotropic loci

Our analyses further support the influence of rare coding variants in *BRCA2* on cancer risk and identified a small set of genes with suggestive associations for future study. Scalable methods for rare variant analyses using GLMMs are now becoming available and continue to be developed. Future studies of larger and more diverse cohorts with WGS data may provide greater insight into the shared genetic architecture of cancer. These studies could investigate the common genetic etiology of cancer by treating cancer as a single homogenous trait, as was done in our analyses. Another option would be to investigate smaller groups of cancers with known genetic correlations as separate traits to look for patterns in associations across broader sets of cancer types. Methodological frameworks to combine results from set-based tests on different phenotypes are currently limited to continuous traits, but work to expand these methods to binary traits is ongoing [104,105]. Finally, incorporating more refined functional annotations when conducting variant set tests could improve statistical power for future rare variant analyses.

Acknowledgements

This work would not have been possible without the support of my thesis committee chair, Dr. Sara Lindström, and my thesis committee. I also collaborated closely with Tongqiu Jia in designing and implementing the exome sequencing data processing, variant annotation, and variant filtering pipelines used in my project. Finally, I wanted to acknowledge other Lindström lab members, Tabitha Harrison and Boya Guo, for their support in discussing ideas and preparing the project.

References

1. World Health Organization (WHO). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death (2020).
2. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
3. Cao, W., Chen, H.-D., Yu, Y.-W., Li, N. & Chen, W.-Q. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. *Chin. Med. J. (Engl.)* **134**, 783–791 (2021).
4. Lortet-Tieulent, J., Georges, D., Bray, F. & Vaccarella, S. Profiling global cancer incidence and mortality by socioeconomic development. *Int. J. Cancer* **147**, 3029–3036 (2020).
5. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **68**, 820–823 (1971).
6. Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643–646 (1986).
7. Malkin, D. *et al.* Germ Line p53 Mutations in a Familial Syndrome of Breast Cancer, Sarcomas, and Other Neoplasms. *Science* **250**, 1233–1238 (1990).
8. Olschwang, S. *et al.* Genetic characterization of the APC locus involved in familial adenomatous polyposis. *Gastroenterology* **101**, 154–160 (1991).
9. Stambolic, V. *et al.* Negative Regulation of PKB/Akt-Dependent Cell Survival by the Tumor Suppressor PTEN. *Cell* **95**, 29–39 (1998).
10. Bell, D. W. *et al.* Heterozygous Germ Line hCHK2 Mutations in Li-Fraumeni Syndrome. *Science* **286**, 2528–2531 (1999).

11. Miki, Y. *et al.* A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1. *Science* **266**, 66–71 (1994).
12. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792 (1995).
13. Rahman, N. *et al.* PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* **39**, 165–167 (2007).
14. Easton, D. F. *et al.* Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *N. Engl. J. Med.* **372**, 2243–2257 (2015).
15. Mucci, L. A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* **315**, 68–76 (2016).
16. Fletcher, O. & Houlston, R. S. Architecture of inherited susceptibility to common cancer. *Nat. Rev. Cancer* **10**, 353–361 (2010).
17. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
18. Jiang, X. *et al.* Shared heritability and functional enrichment across six solid cancers. *Nat. Commun.* **10**, 431 (2019).
19. Lindström, S. *et al.* Quantifying the Genetic Correlation between Multiple Cancer Types. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **26**, 1427–1435 (2017).
20. Rashkin, S. R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat. Commun.* **11**, 4423 (2020).
21. Sampson, J. N. *et al.* Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for 13 Cancer Types. *JNCI J. Natl. Cancer Inst.* **107**, djv279 (2015).

22. Glubb, D. M. *et al.* Cross-Cancer Genome-Wide Association Study of Endometrial Cancer and Epithelial Ovarian Cancer Identifies Genetic Risk Regions Associated with Risk of Both Cancers. *Cancer Epidemiol. Biomarkers Prev.* **30**, 217–228 (2021).
23. Kar, S. P. *et al.* Genome-wide Meta-analyses of Breast, Ovarian and Prostate Cancer Association Studies Identify Multiple New Susceptibility Loci Shared by At Least Two Cancer Types. *Cancer Discov.* **6**, 1052–1067 (2016).
24. Copur, M. & Suresh Manapuram, M. D. Multiple Primary Tumors Over a Lifetime. *Oncology* **33**, (2019).
25. Vogt, A. *et al.* Multiple primary tumours: challenges and approaches, a review. *ESMO Open* **2**, e000172 (2017).
26. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
27. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* 1–7 (2021) doi:10.1038/s41586-021-04103-z.
28. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
29. Gorlov, I. P., Kimmel, M. & Amos, C. I. Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. *Hum. Mol. Genet.* **15**, 1143–1150 (2006).
30. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
31. Hilbers, F. S. *et al.* Rare variants in XRCC2 as breast cancer susceptibility alleles. *J. Med. Genet.* **49**, 618–620 (2012).
32. McGuire Sams, C., Shepp, K., Pugh, J., Bishop, M. R. & Merner, N. D. Rare and potentially pathogenic variants in hydroxycarboxylic acid receptor genes identified in breast cancer cases. *BMC Med. Genomics* **14**, 284 (2021).

33. Liu, Y. *et al.* Rare Variants in Known Susceptibility Loci and Their Contribution to Risk of Lung Cancer. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **13**, 1483–1495 (2018).
34. McWilliams, R. R. *et al.* CDKN2A Germline Rare Coding Variants and Risk of Pancreatic Cancer in Minority Populations. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **27**, 1364–1370 (2018).
35. Permuth, J. B. *et al.* Exome genotyping arrays to identify rare and low frequency variants associated with epithelial ovarian cancer risk. *Hum. Mol. Genet.* **25**, 3600–3612 (2016).
36. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203 (2018).
37. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
38. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat. Genet.* **52**, 242–243 (2020).
39. Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
40. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
41. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
42. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
43. Haworth, S. *et al.* Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
44. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).

45. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
46. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **89**, (2011).
47. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostat. Oxf. Engl.* **13**, 762–775 (2012).
48. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
49. Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *Am. J. Hum. Genet.* **104**, 260–274 (2019).
50. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* **52**, 634–639 (2020).
51. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
52. UK Biobank. Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource. *Information for researchers - Interim Data Release 2015* https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping_qc.pdf (2015).
53. UK Biobank. UK Biobank - Exome Data Release FAQs / December 2020. https://www.ukbiobank.ac.uk/media/cfulxh52/uk-biobank-exome-release-faq_v9-december-2020.pdf (2020).
54. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).

55. UK Biobank. UK Biobank Malignant Cancer Summary Report. https://biobank.ndph.ox.ac.uk/~bbdatan/CancerSummaryReport.html (2022).
56. Wolfe, D., Dudek, S., Ritchie, M. D. & Pendergrass, S. A. Visualizing genomic information across chromosomes with PhenoGram. *BioData Min.* **6**, 18 (2013).
57. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* 1–6 (2022) doi:10.1038/s41586-022-04558-8.
58. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).
59. Cingolani, P. *et al.* Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* **3**, 35 (2012).
60. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
61. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
62. Galinsky, K. J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).
63. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017).
64. Lee, S., Fuchsberger, C., Kim, S. & Scott, L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies. *Biostat. Oxf. Engl.* **17**, 1–15 (2016).
65. Breslow, N. E. & Clayton, D. G. Approximate Inference in Generalized Linear Mixed Models. *J. Am. Stat. Assoc.* **88**, 9–25 (1993).

66. Lee, S. H. & Werf, J. H. van der. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet. Sel. Evol. GSE* **38**, 25–43 (2005).
67. Gilmour, A. R., Thompson, R. & Cullis, B. R. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* **51**, 1440–1450 (1995).
68. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
69. Mi, H. *et al.* PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394–D403 (2021).
70. The Gene Ontology Consortium *et al.* The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
71. Moynahan, M. E., Chiu, J. W., Koller, B. H. & Jasin, M. Brca1 Controls Homology-Directed DNA Repair. *Mol. Cell* **4**, 511–518 (1999).
72. Patel, K. J. *et al.* Involvement of Brca2 in DNA Repair. *Mol. Cell* **1**, 347–357 (1998).
73. Moynahan, M. E., Pierce, A. J. & Jasin, M. BRCA2 Is Required for Homology-Directed Repair of Chromosomal Breaks. *Mol. Cell* **7**, 263–272 (2001).
74. Kerr, P. & Ashworth, A. New complexities for BRCA1 and BRCA2. *Curr. Biol.* **11**, R668–R676 (2001).
75. Welcsh, P. L. & King, M.-C. BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. *Hum. Mol. Genet.* **10**, 705–713 (2001).
76. King, M.-C., Marks, J. H. & Mandell, J. B. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. *Science* **302**, 643–646 (2003).
77. Leongamornlert, D. *et al.* Germline BRCA1 mutations increase prostate cancer risk. *Br. J. Cancer* **106**, 1697–1701 (2012).

78. Kote-Jarai, Z. *et al.* BRCA2 is a moderate penetrance gene contributing to young-onset prostate cancer: implications for genetic testing in prostate cancer patients. *Br. J. Cancer* **105**, 1230–1234 (2011).
79. Ferrone, C. R. *et al.* BRCA Germline Mutations in Jewish Patients With Pancreatic Adenocarcinoma. *J. Clin. Oncol.* **27**, 433–438 (2009).
80. Iqbal, J. *et al.* The incidence of pancreatic cancer in BRCA1 and BRCA2 mutation carriers. *Br. J. Cancer* **107**, 2005–2009 (2012).
81. de Jonge, M. M. *et al.* Endometrial Cancer Risk in Women With Germline BRCA1 or BRCA2 Mutations: Multicenter Cohort Study. *JNCI J. Natl. Cancer Inst.* **113**, 1203–1211 (2021).
82. Hearle, N. *et al.* Contribution of Germline Mutations in BRCA2, P16 INK4A , P14 ARF and P15 to Uveal Melanoma. *Invest. Ophthalmol. Vis. Sci.* **44**, 458–462 (2003).
83. Moran, A. *et al.* Risk of cancer other than breast or ovarian in individuals with BRCA1 and BRCA2 mutations. *Fam. Cancer* **11**, 235–242 (2012).
84. Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat. Genet.* **46**, 736–741 (2014).
85. Lee, K. C., Higgins, H. W. & Qureshi, A. A. Familial risk of melanoma and links with other cancers. *Melanoma Manag.* **2**, 83–89 (2015).
86. Mersch, J. *et al.* Cancers Associated with BRCA1 and BRCA2 Mutations other than Breast and Ovarian. *Cancer* **121**, 269–275 (2015).
87. Gumaste, P. V. *et al.* Skin cancer risk in BRCA1/2 mutation carriers. *Br. J. Dermatol.* **172**, 1498–1506 (2015).
88. Koh, M. J. *et al.* Gastric-type gene expression and phenotype in non-terminal respiratory unit type adenocarcinoma of the lung with invasive mucinous adenocarcinoma morphology. *Histopathology* **76**, 898–905 (2020).

89. Valverde, P., Healy, E., Jackson, I., Rees, J. L. & Thody, A. J. Variants of the melanocyte–stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat. Genet.* **11**, 328–330 (1995).
90. Manganelli, M. *et al.* Behind the Scene: Exploiting MC1R in Skin Cancer Risk and Prevention. *Genes* **12**, 1093 (2021).
91. Amos, C. I. *et al.* Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. *Hum. Mol. Genet.* **20**, 5012–5023 (2011).
92. Williams, P. F., Olsen, C. M., Hayward, N. K. & Whiteman, D. C. Melanocortin 1 receptor and risk of cutaneous melanoma: A meta-analysis and estimates of population burden. *Int. J. Cancer* **129**, 1730–1740 (2011).
93. Tagliabue, E. *et al.* MC1R gene variants and non-melanoma skin cancer: a pooled-analysis from the M-SKIP project. *Br. J. Cancer* **113**, 354–363 (2015).
94. Tagliabue, E. *et al.* MC1R variants as melanoma risk factors independent of at-risk phenotypic characteristics: a pooled analysis from the M-SKIP project. *Cancer Manag. Res.* **10**, 1143–1154 (2018).
95. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
96. Swanson, J. M. The UK Biobank and selection bias. *The Lancet* **380**, 110 (2012).
97. Keyes, K. M. & Westreich, D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet Lond. Engl.* **393**, 1297 (2019).
98. Huang, J. Y. Representativeness Is Not Representative: Addressing Major Inferential Threats in the UK Biobank and Other Big Data Repositories. *Epidemiology* **32**, 189–193 (2021).

99. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026 (2017).

100. Batty, G. D., Gale, C. R., Kivimäki, M., Deary, I. J. & Bell, S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ* **368**, m131 (2020).

101. Stamatakis, E. *et al.* Is Cohort Representativeness Passé? Poststratified Associations of Lifestyle Risk Factors with Mortality in the UK Biobank. *Epidemiol. Camb. Mass* **32**, 179–188 (2021).

102. Qian, Y. *et al.* Identification of pathogenic retrotransposon insertions in cancer predisposition genes. *Cancer Genet.* **216**, 159–169 (2017).

103. Bozsik, A. *et al.* Complex Characterization of Germline Large Genomic Rearrangements of the BRCA1 and BRCA2 Genes in High-Risk Breast Cancer Patients—Novel Variants from a Large National Center. *Int. J. Mol. Sci.* **21**, 4650 (2020).

104. Dutta, D., Scott, L., Boehnke, M. & Lee, S. Multi-SKAT: General framework to test for rare variant association with multiple phenotypes. *Genet. Epidemiol.* **43**, 4 (2019).

105. Dutta, D. *et al.* Meta-MultiSKAT: Multiple phenotype meta-analysis for region-based association test. *Genet. Epidemiol.* **43**, 800 (2019).

Tables and Figures

Table 1. Cancer types represented by the 2,491 cancer-associated SNPs in the GWAS Catalog.

Cancer type	*Percentage of all SNPs*	*N*
Bladder	0.60	15
Breast	20.75	517
Central nervous system	1.81	45
Cervical	1.24	31
Colorectal	9.67	241
Endometrial	1.00	25
Esophageal	1.04	26
Ewing sarcoma	0.24	6
Gastric cancer	0.8	20
Head and neck squamous cell carcinoma	0.12	3
Kidney cancer	1.16	29
Laryngeal cancer	0.16	4
Leukemia	4.74	118
Liver cancer	0.48	12
Lung cancer	6.22	155
Lymphoma	2.29	57
Melanoma	4.34	108
Multiple myeloma	1.45	36
Neuroendocrine tumors	0.08	2
Non-melanoma skin	8.99	224
Pancreatic	1.41	35
Prostate	19.07	475
Pleiotropy	7.79	194
Oral	0.44	11
Osteosarcoma	0.08	2
Ovarian	1.85	46
Testicular	1.08	27
Thyroid	1.00	25
Uveal melanoma	0.08	2

Table 2. Subject demographic characteristics (N = 195,507).

Characteristic	*Overall % (N)*	*Case % (N)*	*Control % (N)*
Sex			
Male	45.2 (88,412)	44.6 (18,570)	45.4 (69,842)
Female	55.8 (107,095)	55.4 (23,100)	54.6 (83,995)
Age at recruitment			
38 – 45	13.1 (25,639)	5.5 (2,280)	15.2 (23,359)
46 – 55	29.4 (57,549)	19.9 (8,277)	32.0 (49,272)
56 – 65	43.1 (84,287)	50.8 (21,159)	41.0 (63,128)
66 – 72	14.4 (28,032)	23.9 (9,954)	11.8 (18,078)
Self-reported background			
Asian	2.15 (4,198)	0.91 (379)	2.48 (3,819)
Black	1.62 (3,158)	0.85 (354)	1.82 (2,804)
Chinese	0.32 (628)	0.15 (61)	0.37 (567)
Do not know	0.04 (71)	0.03 (12)	0.04 (59)
Mixed	0.65 (1,267)	0.38 (159)	0.72 (1,108)
No response	0.11 (209)	0.07 (30)	0.12 (179)
Other	0.97 (1,891)	0.59 (246)	1.07 (1,645)
Prefer not to answer	0.36 (696)	0.34 (140)	0.36 (556)
White	93.80 (183,389)	96.69 (40,289)	93.02 (143,100)

Table 3. Primary cancer site distribution among the 41,670 cancer cases.

Cancer site	*Percentage of all cases*	*N*
Adrenal	0.03	13
Anus	0.21	89
In situ anus	0.03	12
Appendix	0.10	40
Bladder	1.32	551
In situ bladder	0.75	314
Bone	0.12	48
Brain	0.79	328
Breast	16.45	6,853
In situ breast	2.06	859
Cervix	0.94	391
In situ cervix	4.53	1,889
Colorectal	6.38	2,659
In situ colorectal	0.27	113
Endometrium	1.87	779
In situ endometrium	0.43	179
Esophagus	0.86	358
In situ esophagus	0.01	5
Eye	0.17	71
In situ eye	0.19	79
Gallbladder	0.10	41
Head and neck	1.19	495
In situ head and neck	0.02	10
Heart, mediastinum, pleura	0.02	8
Hematopoietic	0.04	15
Hodgkin's lymphoma	0.44	185
Kidney	1.49	619
Larynx	0.29	122
In situ larynx	0.05	19
Liver	0.47	195
Lung	2.72	1,135
In situ lung	0.005	2
Lymphoid leukemia	0.84	352
Melanoma	4.19	1,746

In situ melanoma	1.10	460
Meninges	0.01	6
Mesothelial and connective tissue	0.85	355
Multiple myeloma	0.83	345
Myeloid leukemia	0.56	232
Nasal	0.06	24
Non-Hodgkin's lymphoma	2.58	1,075
Non-melanoma skin	26.98	11,243
In situ non-melanoma skin	0.66	276
Other central nervous system	0.02	9
Other uterine sites	0.14	57
Ovary	1.40	584
Pancreas	0.91	378
Penis	0.06	25
Placenta	0.01	5
Prostate	11.67	4,863
Renal pelvis	0.09	38
Small intestine	0.24	98
Stomach	0.67	280
In situ stomach	0.002	1
Testis	0.75	311
Thymus	0.03	12
Thyroid	0.74	309
Trachea	0.010	4
Ureter	0.06	26
Vagina	0.03	14
Vulva	0.16	66

Table 4. SKAT-O tests results for suggestive genes ($p < 1 \times 10^{-3}$) from the high and moderate impact variant sets.

Gene	Chromosome	SKAT-O *p-value*
MUCL3	6	8.27×10^{-6}
MC1R	16	9.49×10^{-6}
PRDM7	16	4.48×10^{-5}
VARS2	6	9.06×10^{-5}
BRCA2	13	1.46×10^{-4}
COA3	17	1.77×10^{-4}
TBC1D10C	11	3.29×10^{-4}
STUM	1	3.94×10^{-4}
BCAS3	17	4.98×10^{-4}
CALCR	7	5.04×10^{-4}
RBM6	3	5.16×10^{-4}
SMU1	9	7.00×10^{-4}
TMEM145	19	7.94×10^{-4}
NAGS	17	9.17×10^{-4}

Table 5. SKAT-O test results for significant and suggestive genes ($p < 1 \times 10^{-3}$) from the high impact variant sets.

Gene	Chromosome	SKAT-O *p-value*
BRCA2	**13**	**1.09×10^{-6}**
BRCA1	17	6.08×10^{-6}
OR5B21	11	1.40×10^{-4}
RARG	12	2.36×10^{-4}
EDN3	20	3.10×10^{-4}
EIF2A	3	3.91×10^{-4}
KRT75	12	4.49×10^{-4}
AMACR	5	7.11×10^{-4}

*Genes with significant associations are highlighted in bold

Table 6. Single variant test results for significant and suggestive variants ($p < 1 \times 10^{-5}$).

Chromosome	Position	Identifier	Major/Minor allele	Predicted consequence	Gene	*p*-value	MAF
16	**89,919,709**	**rs1805007**	**C/T**	**Missense**	***MC1R***	**1.83×10^{-18}**	**9.62×10^{-2}**
16	89,919,736	rs1805008	C/T	Missense	*MC1R*	6.54×10^{-7}	8.33×10^{-2}
6	32,396,114	rs763452732	C/T	Missense	*BTNL2*	3.71×10^{-6}	3.07×10^{-5}
6	10,410,053	rs1407939076	T/C	Missense	*TFAP2A*	9.10×10^{-6}	3.84×10^{-5}

*Variants with significant associations are highlighted in bold

Table 7. Unconditional and conditional single variant test results for suggestive and significant ($p < 1 \times 10^{-5}$) *MC1R* single variant tests.

Variant identifier	Single variant test type	Odds Ratio	95% CI for Odds Ratio	*p*-value
rs1805007	Unconditional	1.12	1.09 – 1.15	1.83×10^{-18}
rs1805007	Conditioning on rs1805008	1.14	1.11 – 1.17	8.63×10^{-21}
rs1805008	Unconditional	1.07	1.04 – 1.10	6.54×10^{-7}
rs1805008	Conditioning on rs1805007	1.09	1.06 – 1.12	2.60×10^{-9}

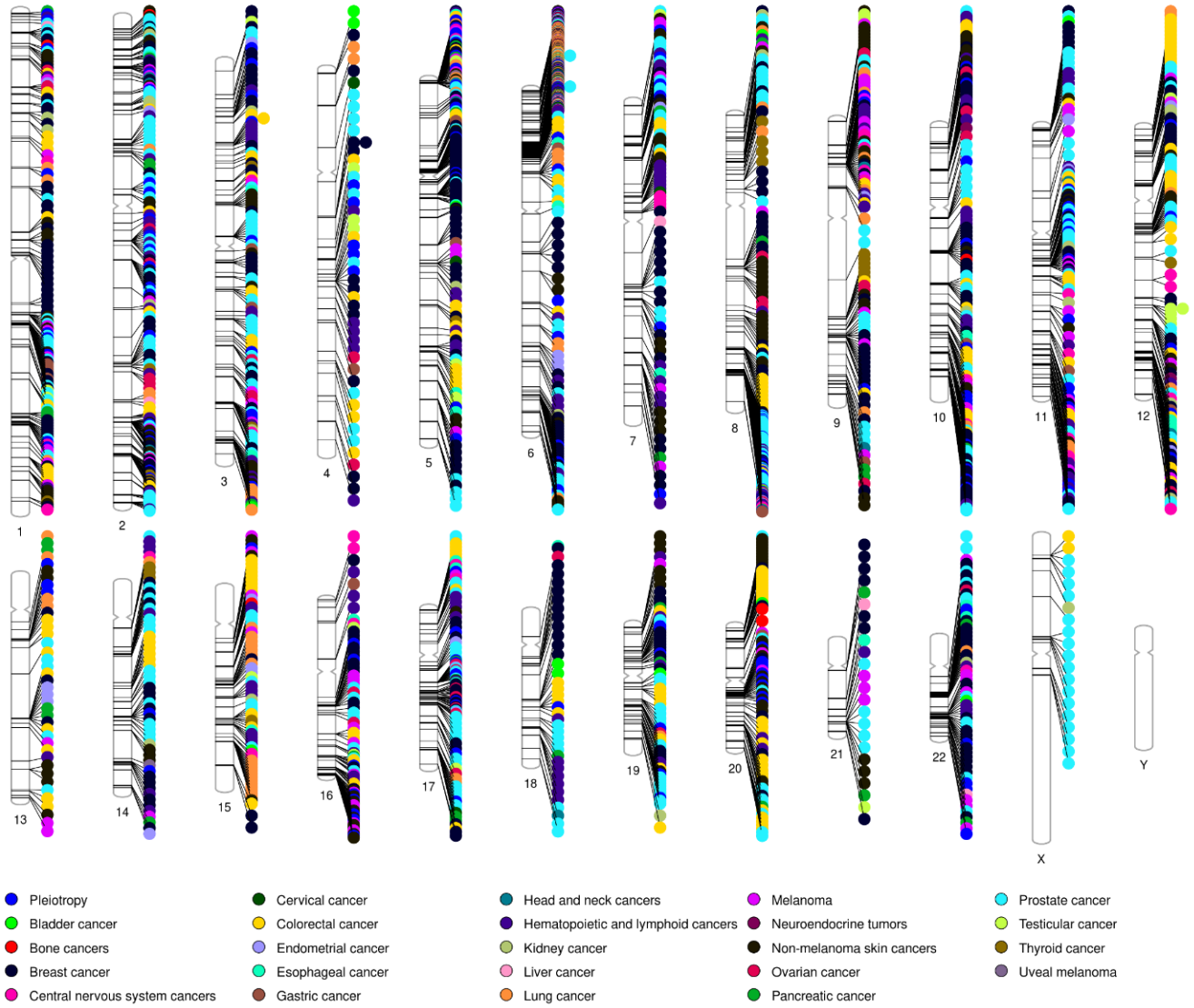


Figure 1. Genomic locations of cancer associated SNPs in GWAS Catalog. The GRCh38 genomic locations of the 2,491 SNPs with reported genome-wide significant cancer associations in the GWAS Catalog are shown in this diagram. Genomic positions are represented by the black lines on each chromosome. The cancer type for the reported SNP association is represented by the colored circles to the right of the chromosomes on the diagram. SNPs with reported pleiotropic associations or associations to at least two cancer types are labeled in dark blue.

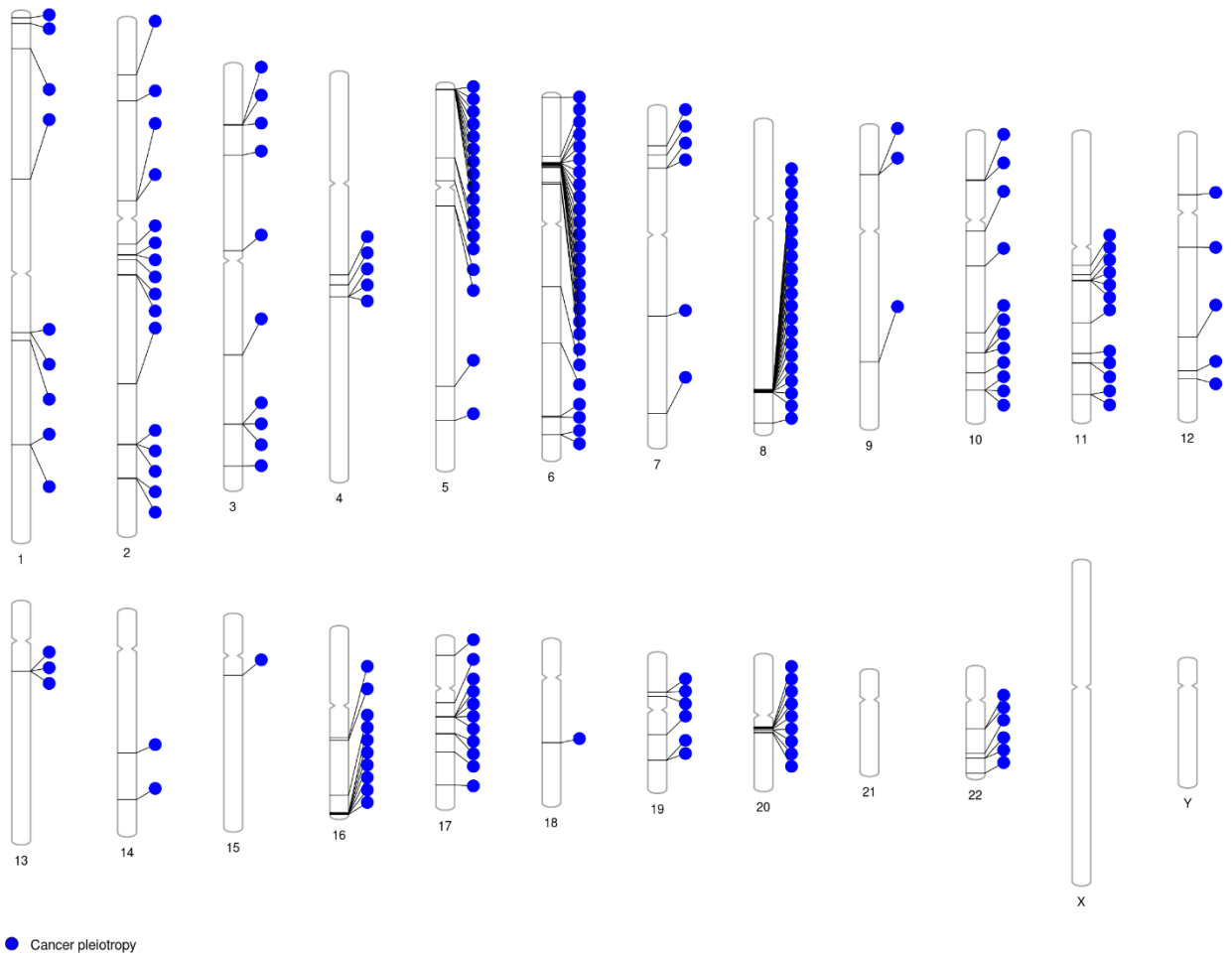


Figure 2. Genomic locations of pleiotropic cancer associated SNPs in GWAS Catalog. The GRCh38 genomic locations of the 194 SNPs with reported genome-wide significant pleiotropic cancer associations in the GWAS Catalog are shown in this diagram. Genomic positions are represented by the black lines on each chromosome.

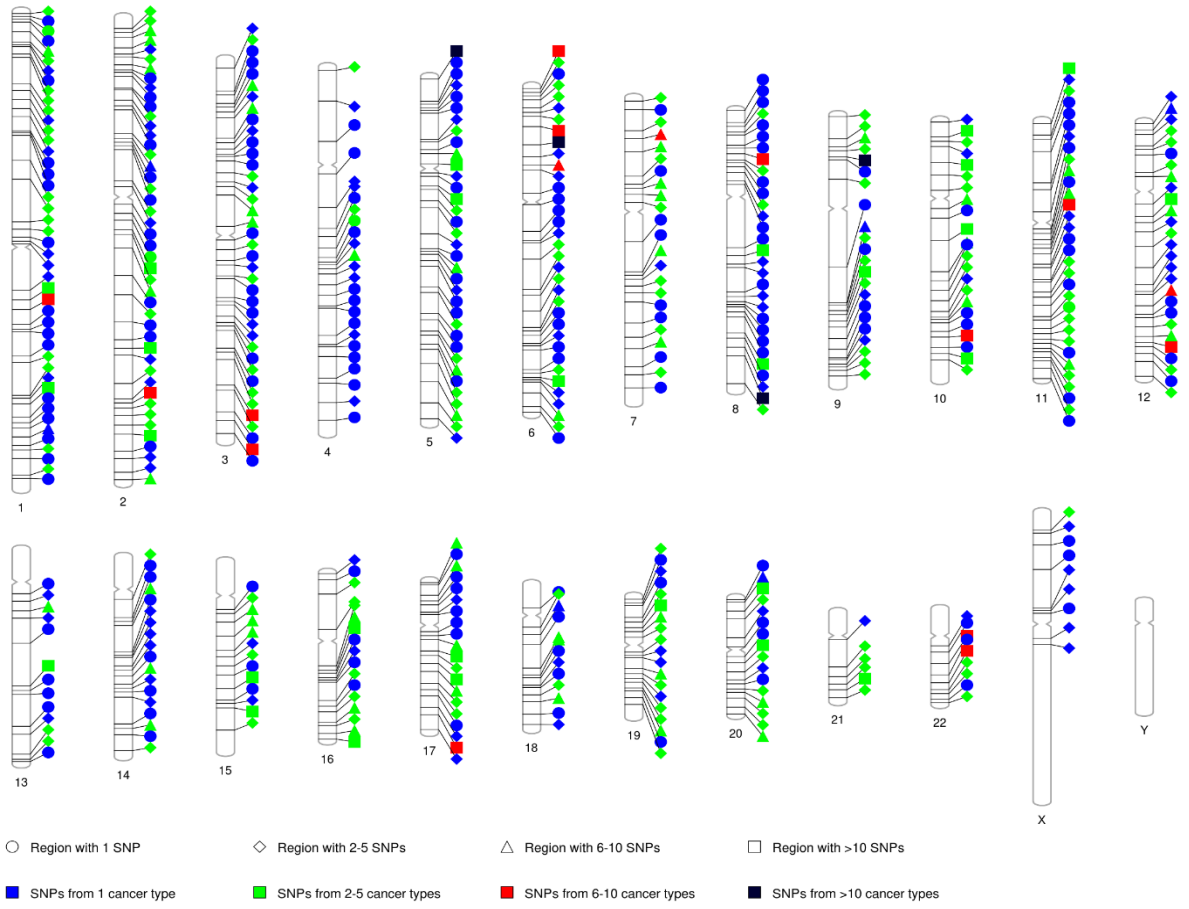


Figure 3. Genomic locations of 534 cancer GWAS susceptibility regions. The midpoints for non-overlapping genomic regions containing the 2,491 SNPs with reported genome-wide cancer associations in the GWAS Catalog are shown in this diagram. These regions are plotted using GRCh38 genomic coordinates. Regions were created by putting each cancer-associated SNP at the midpoint of a 1 Mb window and combining overlapping windows. Region label shapes show the number of SNPs included in that cancer GWAS susceptibility region. The colors of region labels indicate the number of SNP associations to different cancer types contained within that cancer GWAS susceptibility region.

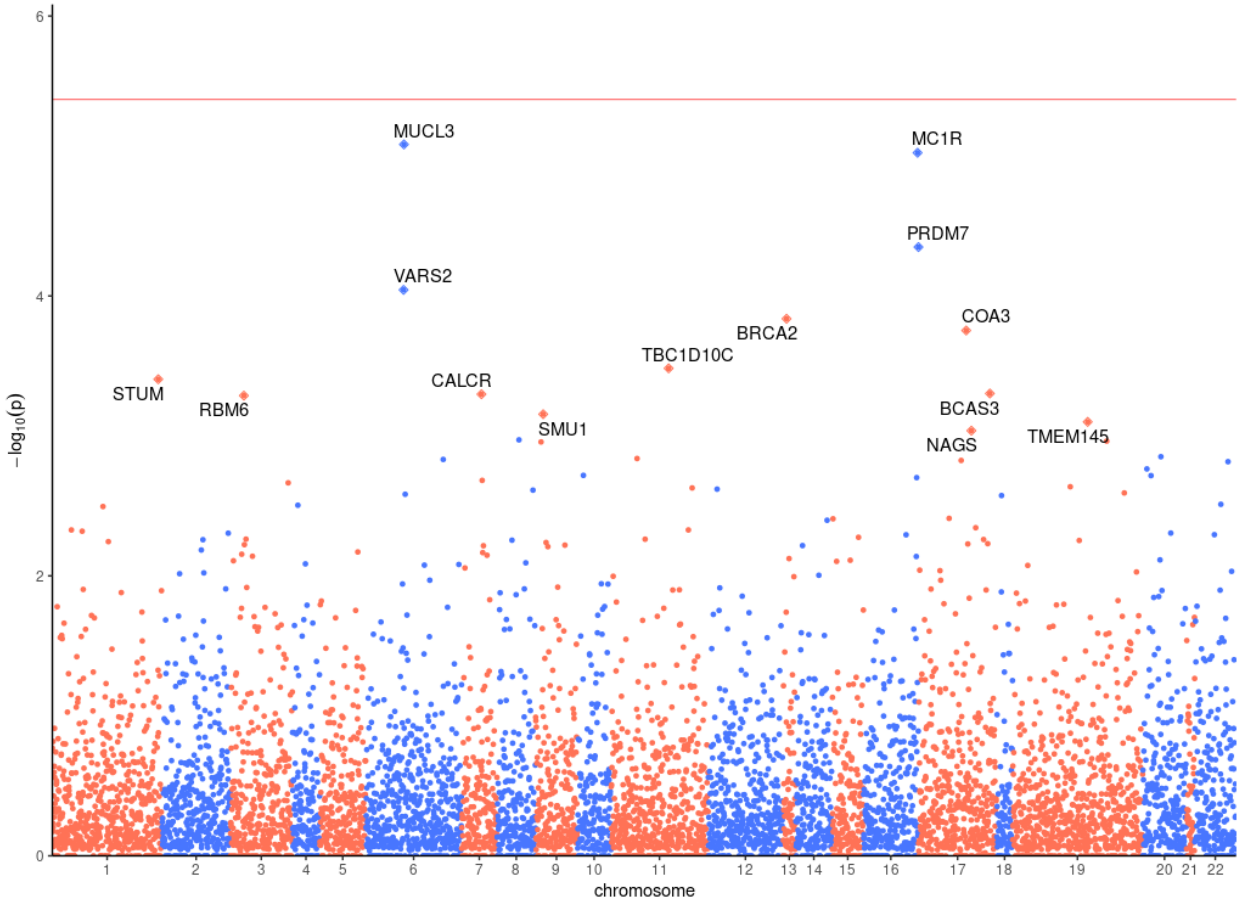


Figure 4. Manhattan plot of SAIGE-GENE SKAT-O test results for gene-based variant sets with high and moderate impact variants. The significance threshold on the plot represents a Bonferroni correction for 12,643 tests ($\alpha = 0.05/12{,}643 = 3.95 \times 10^{-6}$). All genes with SKAT-O $p < 1 \times 10^{-3}$ are highlighted and labeled on the plot.

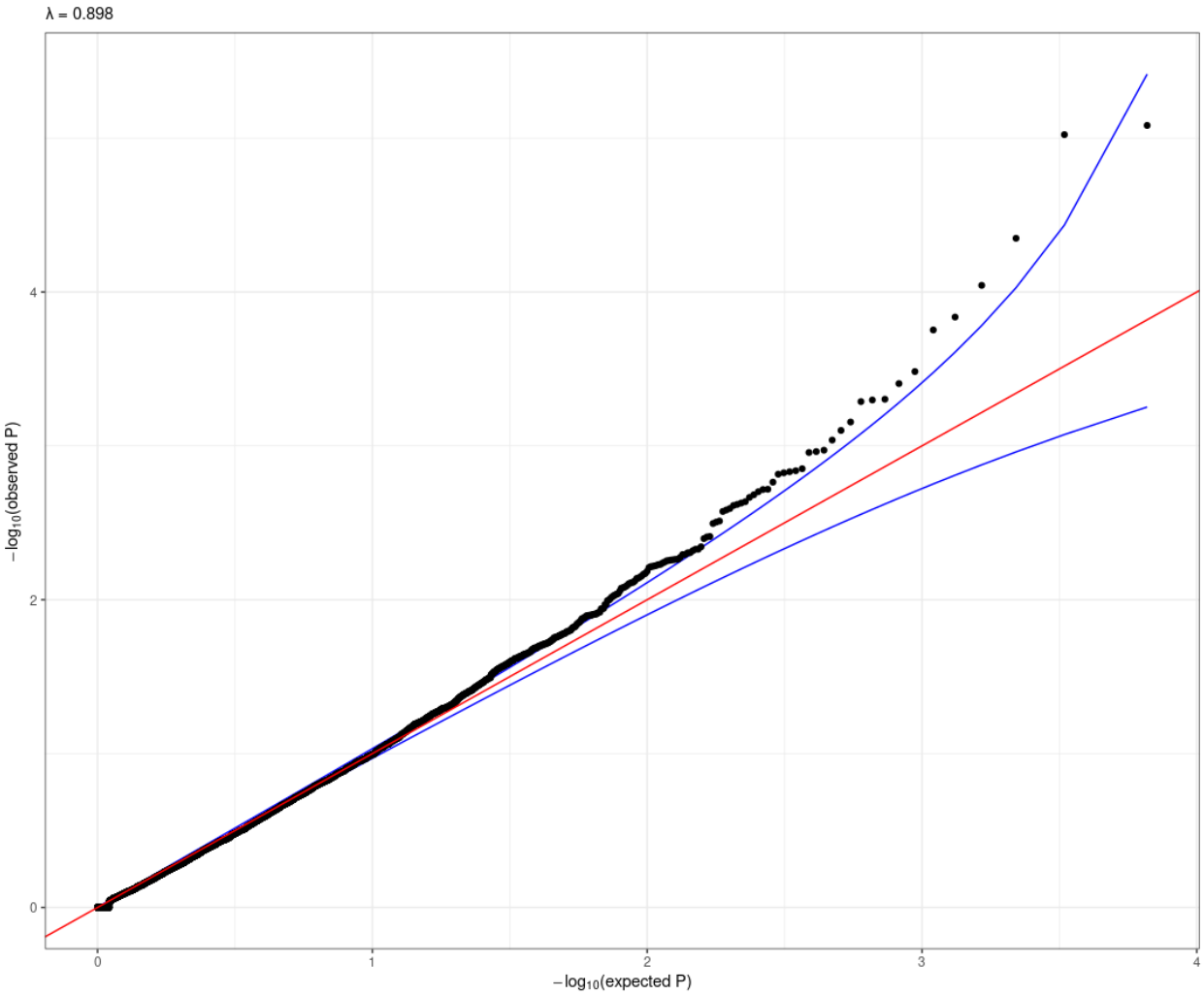


Figure 5. Quantile-Quantile (Q-Q) plot of SAIGE-GENE SKAT-O test results for gene-based variant sets with high and moderate impact variants. The red line indicates the equivalence of the observed SKAT-O and expected *p*-values under the null distribution. The blue lines represent 95% confidence intervals for equivalence under a null *Beta* distribution.

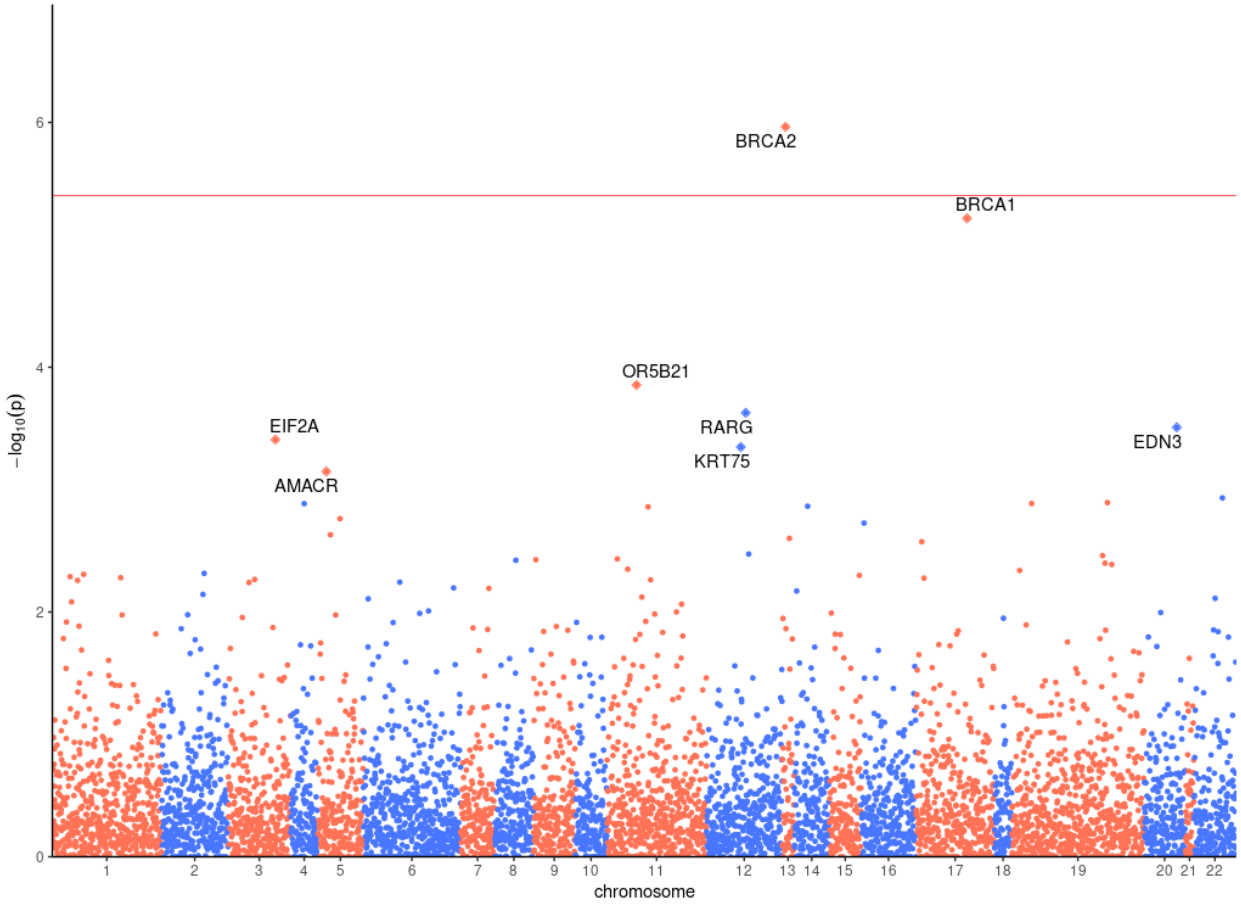


Figure 6. Manhattan plot of SAIGE-GENE SKAT-O test results for gene-based variant sets with high impact variants. The significance threshold on the plot represents a Bonferroni correction for 12,643 tests ($\alpha = 0.05/12{,}643 = 3.95 \times 10^{-6}$). All genes with SKAT-O $p < 1 \times 10^{-3}$ are highlighted and labeled on the plot.

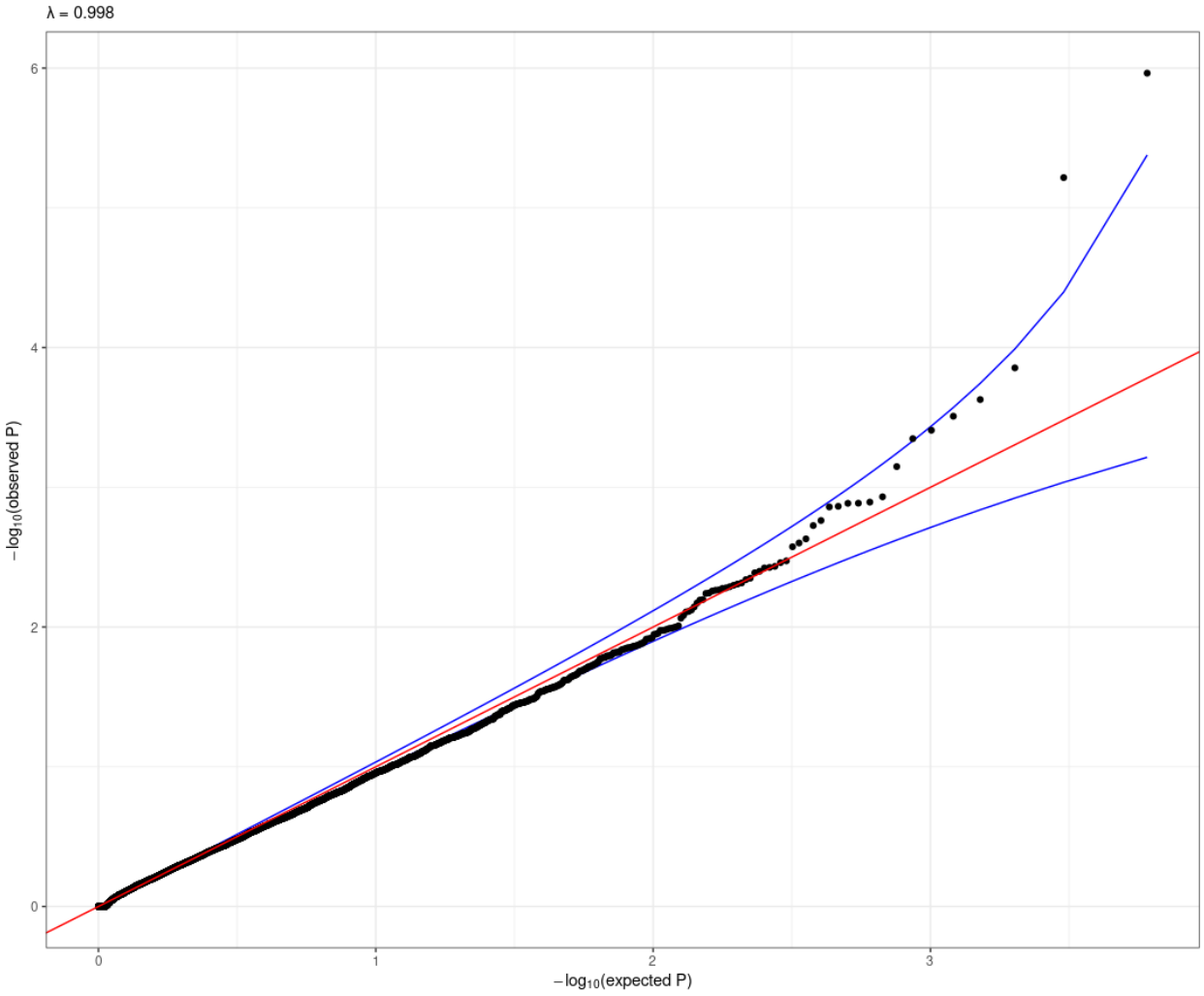


Figure 7. Quantile-Quantile (Q-Q) plot of SAIGE-GENE SKAT-O test results for gene-based variant sets with high impact variants. The red line indicates the equivalence of the observed SKAT-O and expected *p*-values under the null distribution. The blue lines represent 95% confidence intervals for equivalence under a null *Beta* distribution.

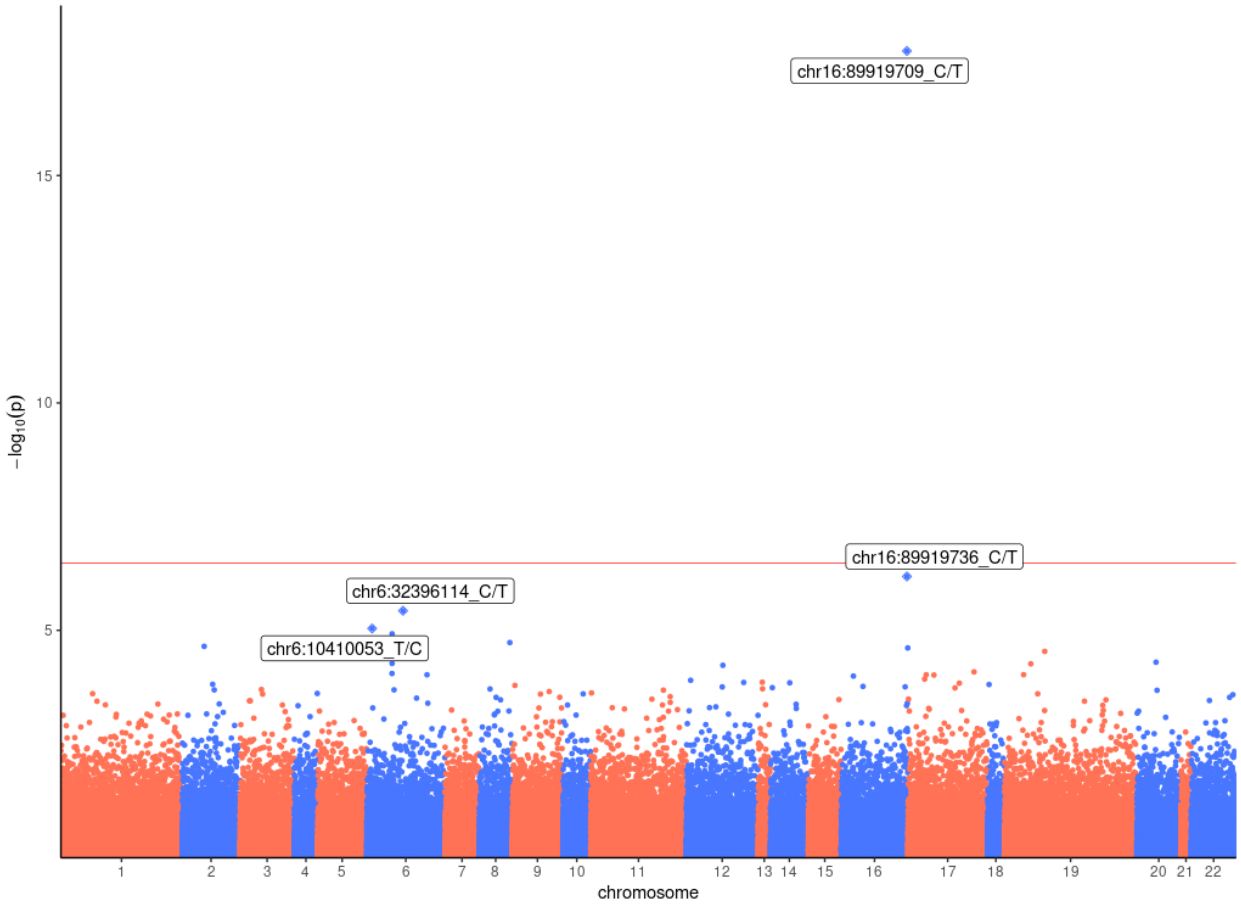


Figure 8. Manhattan plot of SAIGE single variant test results for variants with predicted high and moderate impacts. Ultra-rare variants (MAC ≤ 10) were not included in the single variant tests. The significance threshold on the plot represents a Bonferroni correction for 148,938 tests ($\alpha = 0.05/148{,}938 = 3.36 \times 10^{-7}$). All genes with SKAT-O $p < 1 \times 10^{-5}$ are highlighted and labeled on the plot.

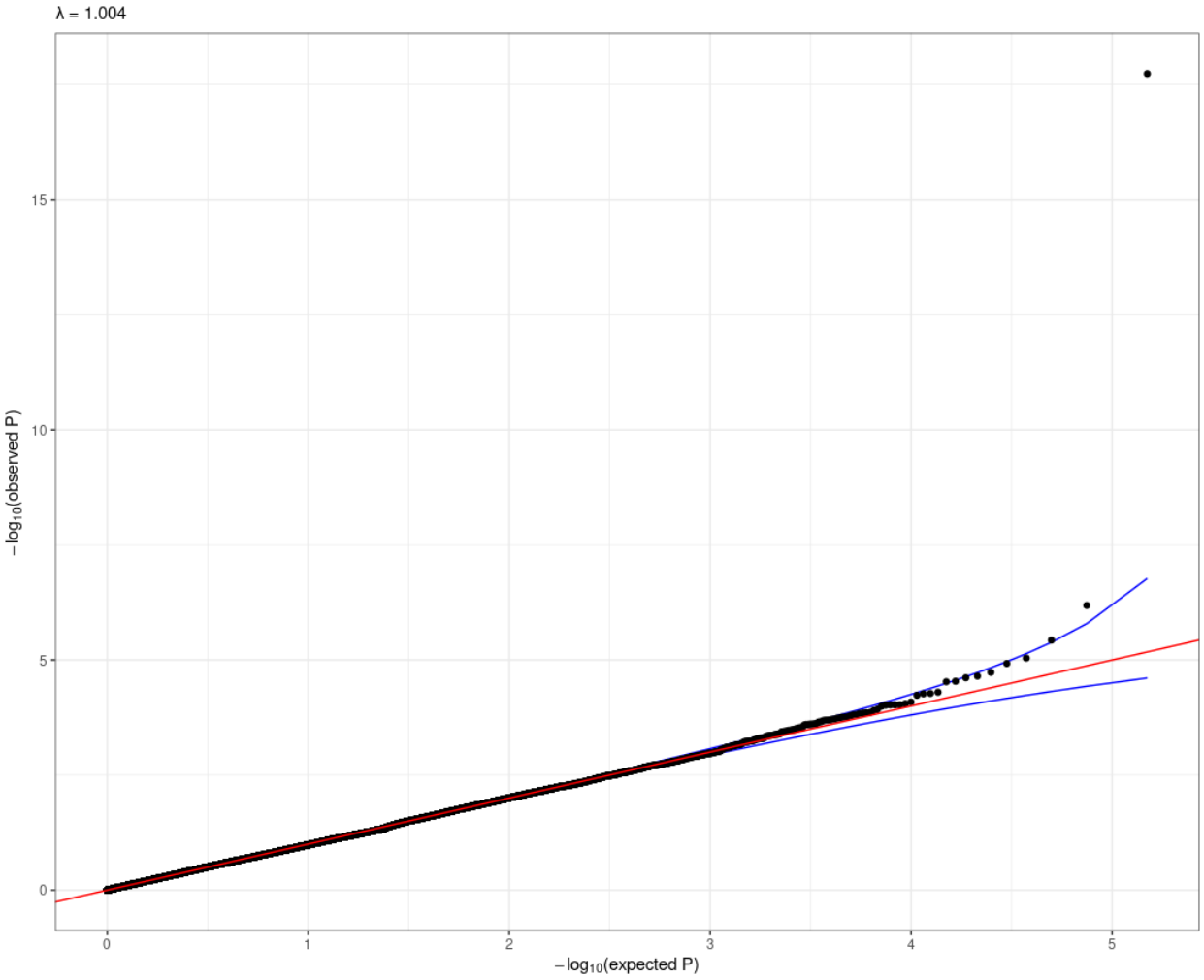
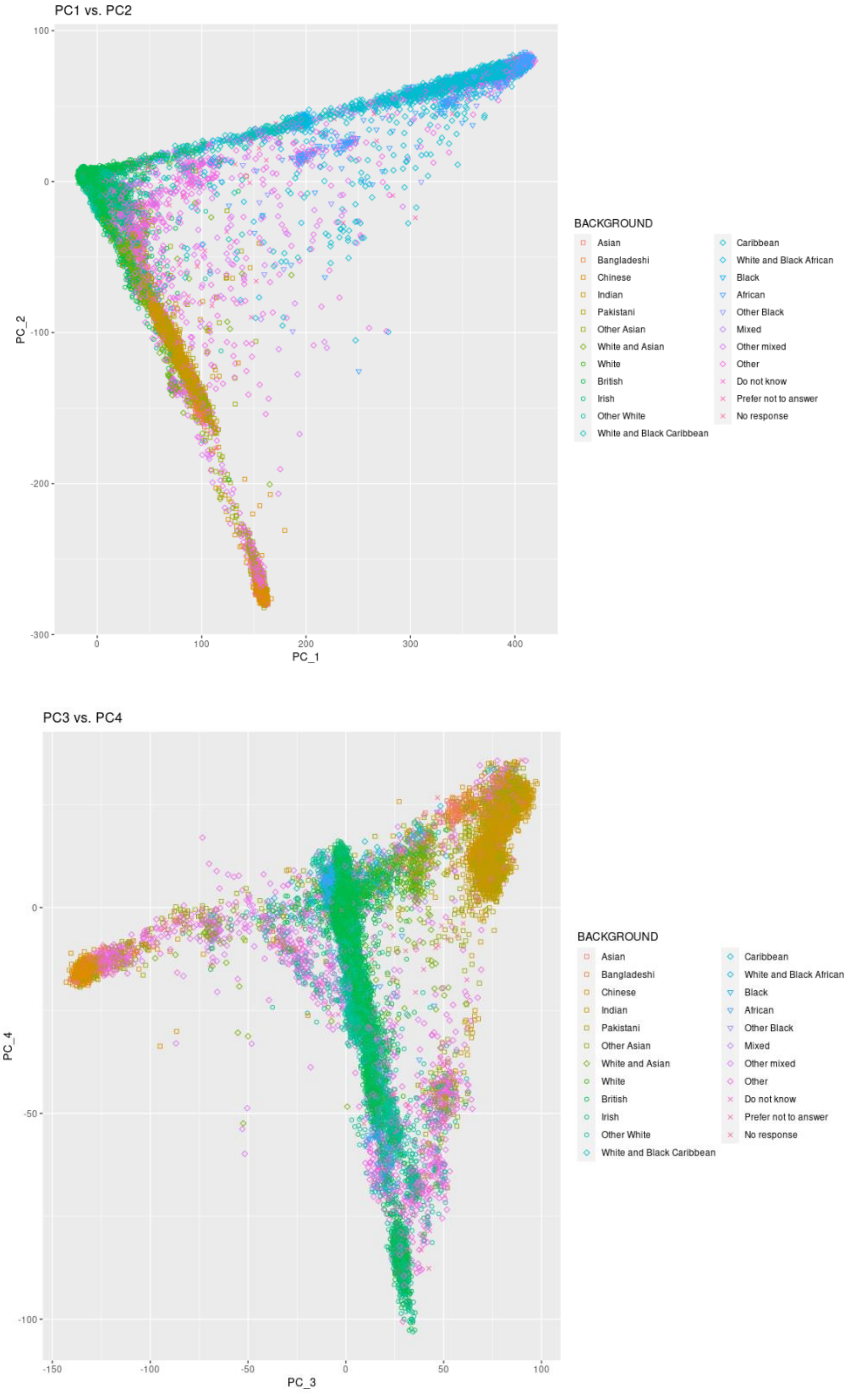


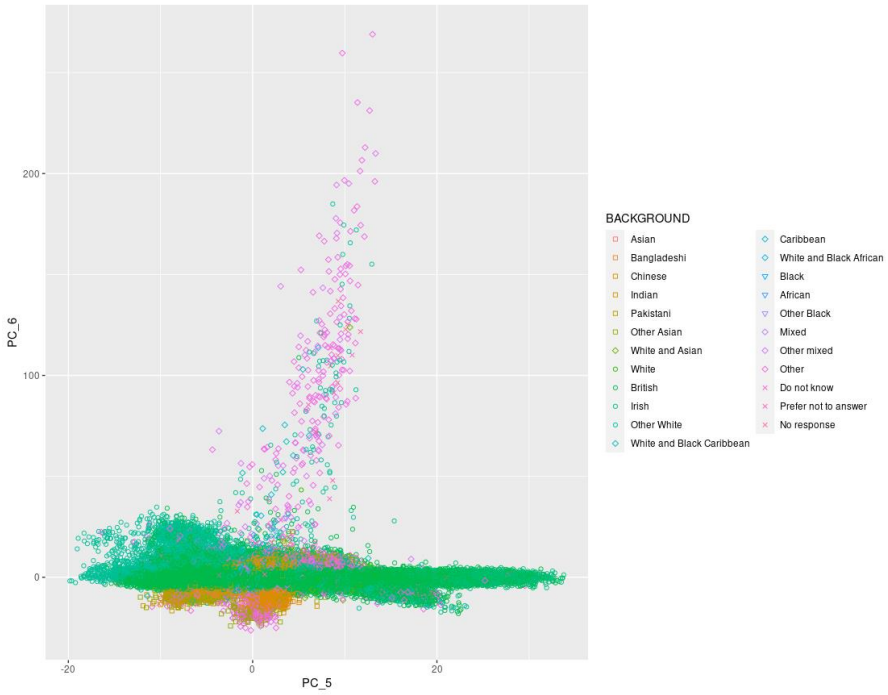
Figure 9. Quantile-Quantile (Q-Q) plot of SAIGE single variant test results for variants with predicted high and moderate impact variants. The red line indicates the equivalence of the observed and expected *p*-values under the null distribution. The blue lines represent 95% confidence intervals for equivalence under a null *Beta* distribution.

Supplementary Figures

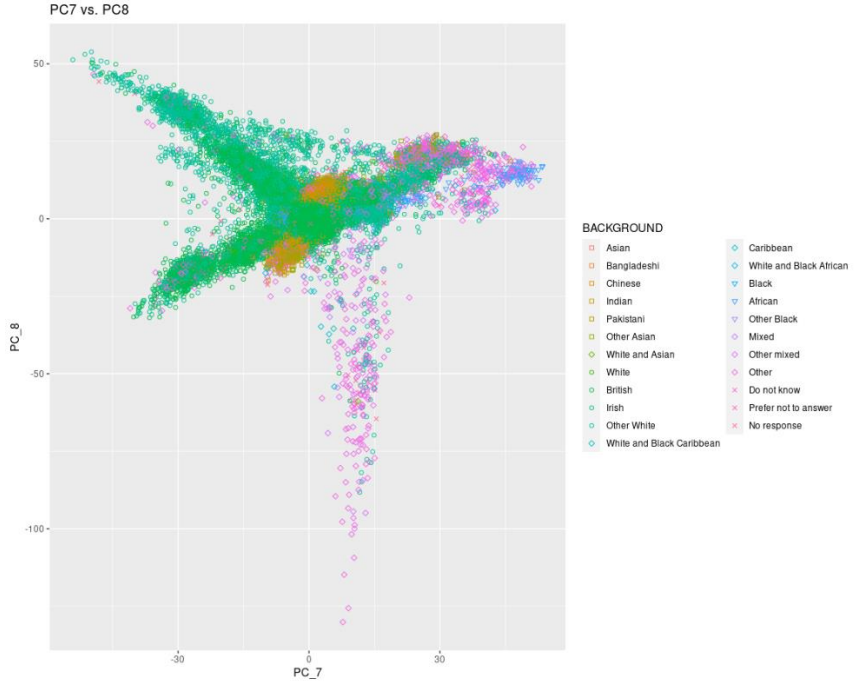
Supplementary figures 1-10. Pairwise principal component plots. Plots of the pairwise principal components for the 195,507 UKBB participants included in the analyses. An individual's self-reported ethnic background is indicated by the color and shape of their point.

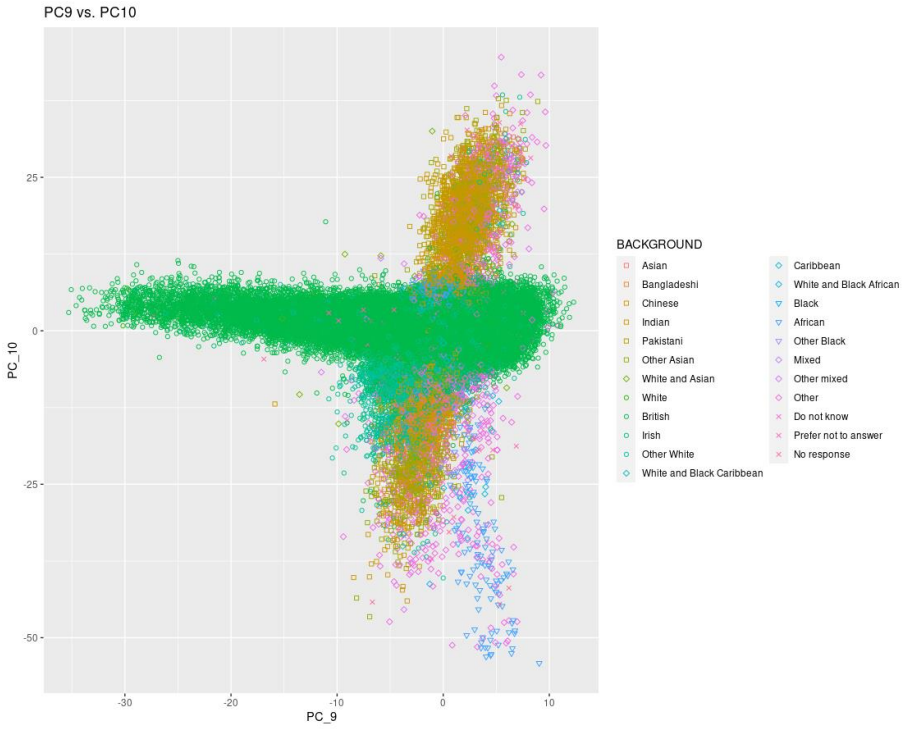


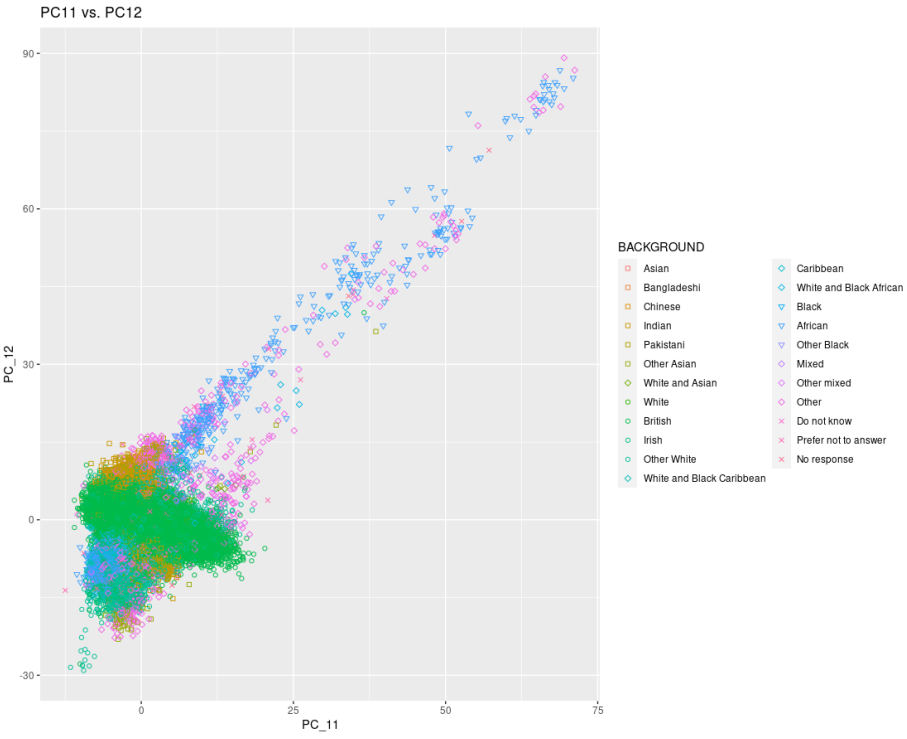
PC5 vs. PC6

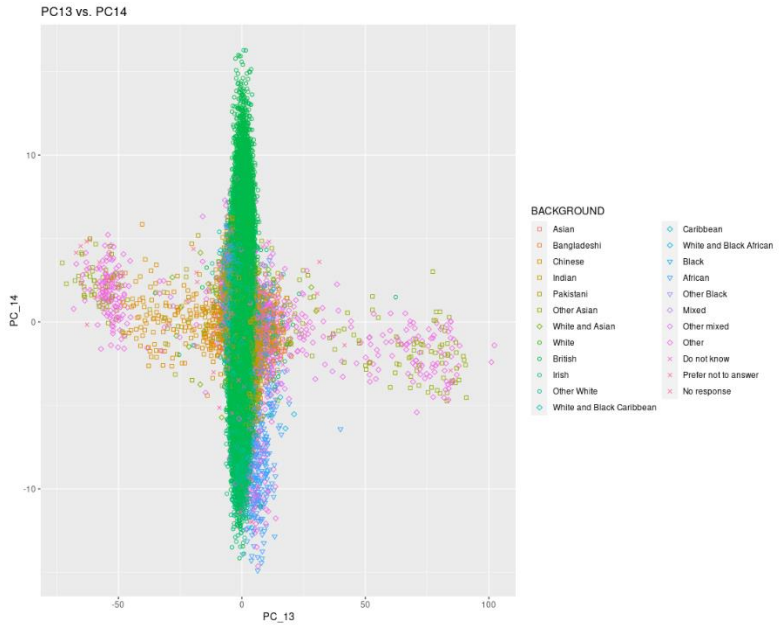


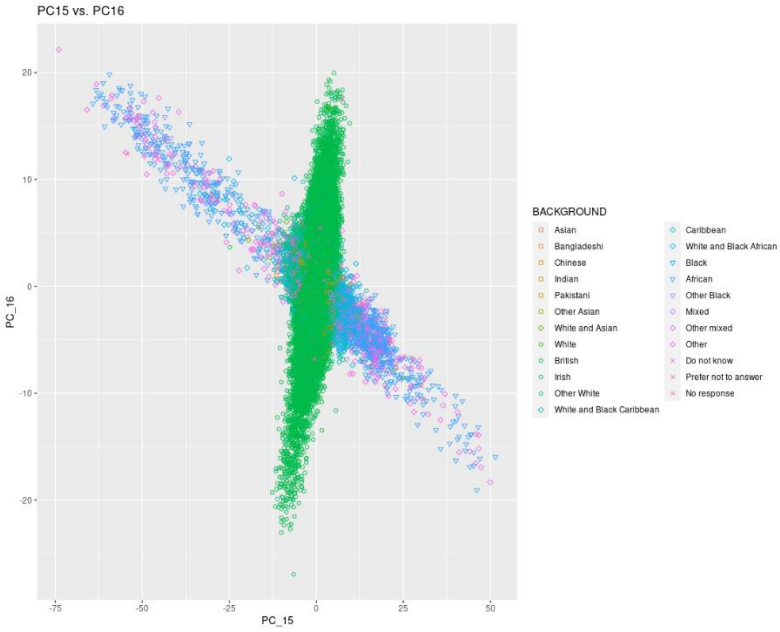
PC7 vs. PC8

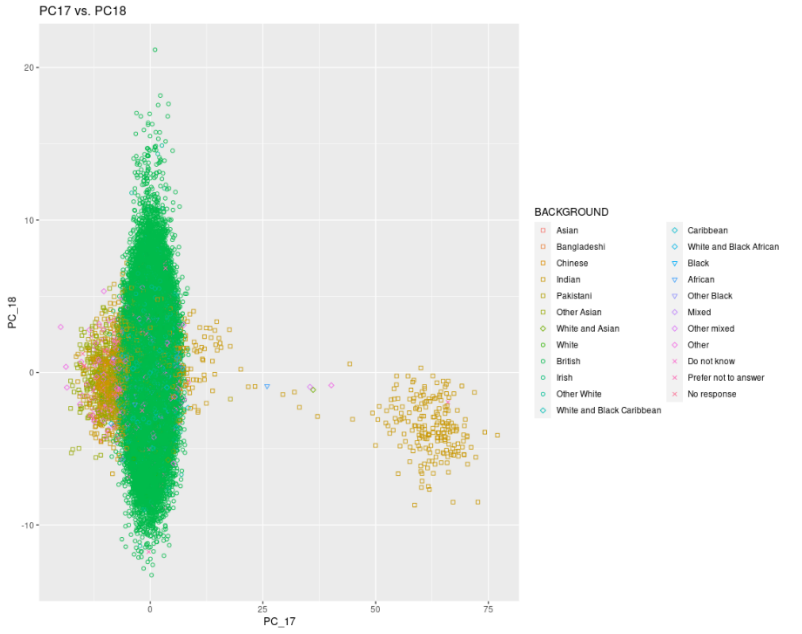


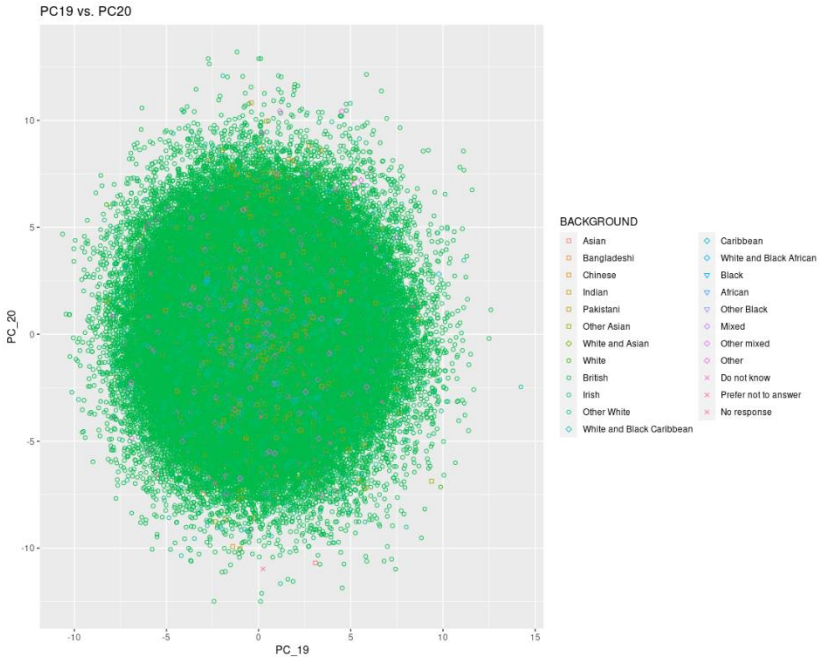

PC9 vs. PC10
PC_10
PC_9
25
0
-25
-50
-30
-20
-10
0
10
BACKGROUND
Asian
Bangladeshi
Chinese
Indian
Pakistani
Other Asian
White and Asian
White
British
Irish
Other White
White and Black Caribbean
Caribbean
White and Black African
Black
African
Other Black
Mixed
Other mixed
Other
Do not know
Prefer not to answer
No response

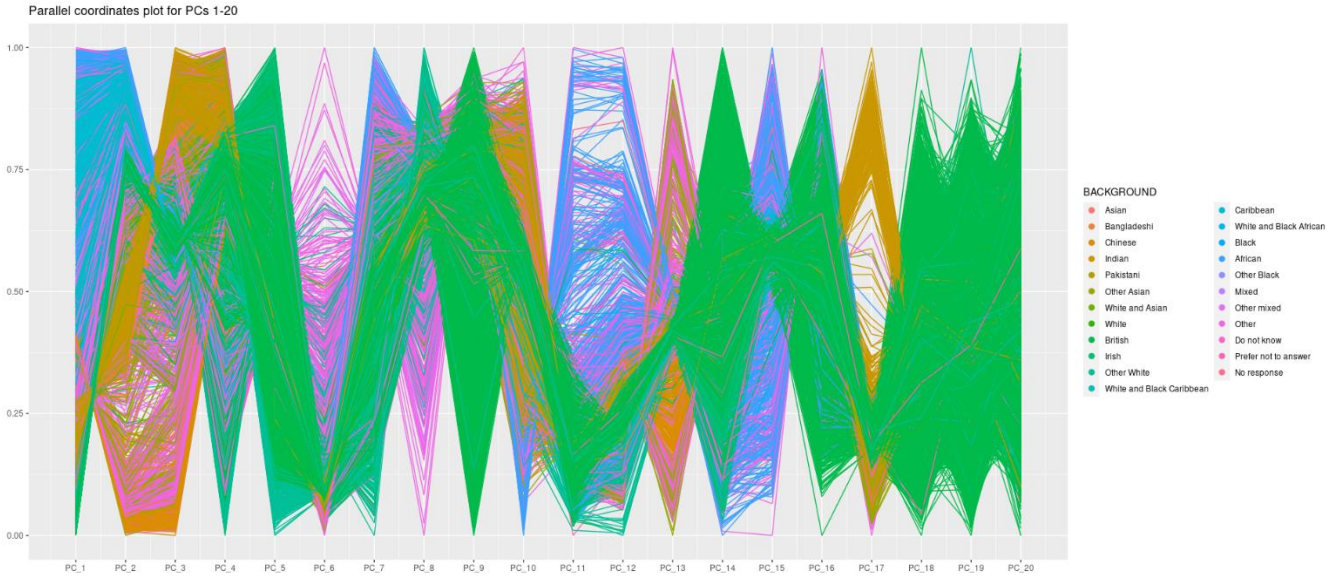

PC11 vs. PC12
PC_12
PC_11
90
60
30
0
-30
0
25
50
75
BACKGROUND
Asian
Bangladeshi
Chinese
Indian
Pakistani
Other Asian
White and Asian
White
British
Irish
Other White
White and Black Caribbean
Caribbean
White and Black African
Black
African
Other Black
Mixed
Other mixed
Other
Do not know
Prefer not to answer
No response


PC13 vs. PC14
PC_14
PC_13
10
0
-10
-50
0
50
100
BACKGROUND
Asian
Bangladeshi
Chinese
Indian
Pakistani
Other Asian
White and Asian
White
British
Irish
Other White
White and Black Caribbean
Caribbean
White and Black African
Black
African
Other Black
Mixed
Other mixed
Other
Do not know
Prefer not to answer
No response


PC15 vs. PC16
PC_16
PC_15
20
10
0
-10
-20
-75
-50
-25
0
25
50
BACKGROUND
Asian
Bangladeshi
Chinese
Indian
Pakistani
Other Asian
White and Asian
White
British
Irish
Other White
White and Black Caribbean
Caribbean
White and Black African
Black
African
Other Black
Mixed
Other mixed
Other
Do not know
Prefer not to answer
No response


PC17 vs. PC18
PC_18
PC_17
20
10
0
-10
0
25
50
75
BACKGROUND
Asian
Bangladeshi
Chinese
Indian
Pakistani
Other Asian
White and Asian
White
British
Irish
Other White
White and Black Caribbean
Caribbean
White and Black African
Black
African
Other Black
Mixed
Other mixed
Other
Do not know
Prefer not to answer
No response


PC19 vs. PC20
PC_20
PC_19
10
5
0
-5
-10
-10
-5
0
5
10
15
BACKGROUND
Asian
Bangladeshi
Chinese
Indian
Pakistani
Other Asian
White and Asian
White
British
Irish
Other White
White and Black Caribbean
Caribbean
White and Black African
Black
African
Other Black
Mixed
Other mixed
Other
Do not know
Prefer not to answer
No response



Supplementary Figure 11. Principal component parallel coordinates plot. A parallel coordinates plot of the first 20 PCs for the 195,507 UKBB participants included in the analyses. An individual's self-reported ethnic background is indicated by the color of their line.