Polygenic background modifies penetrance of monogenic variants conferring risk for coronary artery disease, breast cancer, or colorectal cancer

Supplementary Appendix

Table of Contents

Supplementary methods

Study populations

The study populations consisted of three cohorts. First, a coronary artery disease case-control cohort of 12,879 participants was selected from the UK biobank and underwent exome sequencing at the Broad Institute of MIT and Harvard. Coronary artery disease cases were defined centrally based on self-report at enrollment, hospitalization records, or death registry records (http://biobank.ndph.ox.ac.uk/showcase/showcase/docs/alg_outcome_mi.pdf). Controls included participants free of any documented history of coronary artery disease.

Second, a cohort consisting of 19,264 individuals who underwent clinical grade genetic testing for Hereditary Breast and Ovarian Cancer Syndrome (HBOC) at a commercial testing laboratory (Color Genomics; Burlingame, CA) was also used. Breast cancer case ascertainment was based on self-report at the time of enrollment.[1]

Third, a subset of the UK biobank consisting of 49,738 participants that underwent exome sequencing at Regeneron Genetics Center was used.[2] There was no overlap between this cohort and the 12,879 case-control cohort. Extensive clinical data including diagnosis of prevalent and incident cardiovascular disease and cancer are available on all participants. Coronary artery disease was defined based on self-report of “heart attack/myocardial infarction”, hospitalization records confirming a diagnosis of acute myocardial infarction or ischemic heart disease, coronary revascularization procedures (coronary artery bypass graft surgery or percutaneous angioplasty/stent placement), or death register indicating ischemic heart disease or myocardial infarction as a cause of death. Breast and colorectal cancer were each defined based on self-report of the diagnosis, hospitalization records, cancer register data specifying type of cancer, and death register. For each of the three diseases, we considered the earliest date at which the diagnosis was ascertained as the diagnosis date. All participants diagnosed at dates prior to enrollment in the UK biobank were considered prevalent at baseline, while participants diagnosed after enrollment were considered incident (Table S2).

Informed consent was obtained from all participants. Analysis of UK Biobank data was performed under application number 7089 and approved by the Partners Healthcare institutional review board. The commercial testing laboratory cohort was approved by the Western Institutional Review Board (protocol number 20150716).

Gene sequencing

Whole exome sequencing on 12,909 samples from the coronary artery disease case-control cohort derived from the UK Biobank was performed at the Broad Institute of MIT and Harvard (Cambridge, MA) as previously described.[3] Briefly, libraries were constructed and sequenced on an Illumina HiSeq sequencing using 151 bp pair-end reads.[4] An Illumina Nextera Exome Kit was used for in-solution hybrid selection. Sequencing reads were aligned to the human reference genome build GRCh37.p13 using the Burrows-Wheeler Aligner algorithm,[5] and aligned non-duplicate reads were locally realigned and base quantiles were recalibrated using the Genome Analysis Toolkit software.[6,7] Variants were jointly called using the HaplotypeCaller module of the Genome Analysis Toolkit. We removed samples with contamination >10% (n=0), samples with < 80% of target bases at 20X coverage (n=0), putative sex chromosome aneuploidy (n=17), outliers for heterozygozity (n=4), genotype call rate <95% (n=6), and samples for which there is was no genotyping array data (n=3). Variants from the remaining 12,879 samples were carried forward for further analysis. The mean target coverage was 75X, and 91.1% of target bases were captured at >20X sequencing depth.

For the 19,264 samples from the breast cancer case-control cohort, target enrichment sequencing was performed at the laboratory of Color Genomics (Burlingame, CA) as previously described.[1] The Color Genomics laboratory is in compliance with Clinical Laboratory Improvement Amendments (number 05D2081492) and College of American Pathologists (number 8975161). In brief, sequencing reads were aligned to the human reference genome

build GRCh37.p12 using the Burrows-Wheeler Aligner algorithm, and duplicated and low-quality reads were discarded. Variants were then jointly called using the HaplotypeCaller module of GATK3.4, an internally developed algorithm using SAMtools version 1.8, and dedicated algorithms based on read depth (CNVkit version 0.8.5), paired reads, and split reads (LUMPY version 0.2.13, in-house developed algorithms). A no template control and two positive controls containing a set of known variants were included in every batch of samples. Strict coverage requirements (20 unique reads for each base) were used, and median coverage ranged between 200 and 300X.

Finally, whole-exome sequencing of 49,960 UK Biobank participants was performed at the Regeneron Genetics Center as previously described,[2] and sequencing reads were aligned to the human reference genome build GRCh38 using the Burrows-Wheeler Aligner algorithm. Coverage exceeded 20X at 94.6% of sites on average. Variant calls through two separate pipelines, an "SPB pipeline" that used WeCall (GenomicsPLC) and GLnexus software and a functional equivalence (FE) pipeline, were made available by the UK Biobank for 49,960 samples.[2,8,9] We included variants from the FE pipeline that were also present in the SPB pipeline. We excluded 222 samples for which there were no genotyping data available (n=51) or that failed additional sample quality control using genotyping data: heterozygous missingness outlier (n=112), putative sex chromosome aneuploidy (n=56) and discordance between reported and genetic sex (n=20). The remaining variants on 49,738 participants were carried forward for further analysis. We converted PLINK format to VCF and performed a liftover from GRCh38 to GRCh37.p13.

Variant quality control

In the 12,879 samples from the coronary artery disease case-control cohort, the analysis was limited to the protein-coding regions and canonical splice sites of three familial hypercholesterolemia genes (*LDLR*, *APOB* and *PCSK9*). We then filtered the observed variants

to a candidate list of variants that excludes synonymous variants or variants present at allele frequency of >0.005 in each racial subpopulation of the gnoMAD Genome Aggregation Database, a publicly available population allele frequency database of 141,456 human exomes and genomes.[10]

The variant quality control for the 19,264 samples from the breast cancer case-control cohort was described previously.[1] Briefly, the analysis was limited to rare and high quality variants in the protein-coding regions and canonical splice sites of *BRCA1* and *BRCA2* genes.

In the 49,738 participants form the UK Biobank cohort, variant quality control and classification was previously described in detail.[11] In brief, the analysis was limited to the protein-coding regions and canonical splice sites of 9 genes for any of the three genomic conditions: familial hypercholesterolemia (*LDLR*, *APOB* and *PCSK9*), hereditary breast and ovarian cancer syndrome (*BRCA1* and *BRCA2*), and Lynch syndrome (*MLH1*, *MSH2*, *MSH6* and *PMS2*). We filtered the observed variants to a candidate list of variants that excludes synonymous variants or variants present at allele frequency of >0.005 in any racial subpopulation of the gnomAD Genome Aggregation Database.[10] We also performed additional variant quality control filters excluding variants that fall in low complexity regions, variants that fall in regions with segmental duplications, or variants that do not pass the threshold for the random forest algorithm of gnomAD.[10,12]

Variant classification

For the 12,879 coronary artery disease case-control cohort and the 49,738 UK biobank cohort, candidate variants were filtered to select variants meeting clinical criteria of pathogenicity (pathogenic or likely pathogenic) based on American College of Medical Genetics and Genomics (ACMG)/Association of Molecular Pathology (AMP) criteria,[13] by an American Board of Genetics and Genomics (AMBGG)-certified clinical geneticist, blinded to the phenotype of the participants, at the Partners HealthCare Laboratory of Molecular Medicine (Boston, MA).

In summary, the ACMG/AMP criteria for classifying pathogenic variants look at the effect of the variant on the gene, the previous reports of pathogenicity of the variant, functional studies supporting the damaging effect of the gene, and the prevalence of the variant in cohorts of cases with the disease and controls.[13]

Similarly, in the 19,264 breast cancer case-control cohort from Color Genomics, variant classification was reviewed and signed out by an American Board of Genetics and Genomics (AMBGG)-certified clinical geneticist following criteria for pathogenicity for hereditary breast and ovarian cancer syndrome based on ACMG/AMP criteria.[13]

Polygenic score derivation and ancestry correction

We used three previously validated polygenic scores for coronary artery disease, breast cancer and colorectal cancer containing 6,630,150, 3820, and 95 variants, respectively.[14–16]

Imputed genotype array data available through the UK Biobank was used to calculate the three polygenic scores in all the UK Biobank participants (N= 486,477) as previously described.[3] To minimize confounding caused by population stratification, we fit a linear regression model using the first four principal components of ancestry to predict each of the three polygenic scores. We then used the residuals from these models as ancestry-corrected polygenic score and created reference distributions for each ancestry-corrected score (Fig. S1 in the Supplementary Appendix), as described previously.[17] We determined the percentiles for each individual based on these reference distributions.

In the cohort of 19,264 participants from Color Genomics, low-coverage whole genome sequencing to a minimum depth of 0.2x was performed and variants were imputed for calculation of the polygenic score as previously described.[18] Ancestry-corrected polygenic scores were similarly derived to minimize confounding: we fit a linear regression model that uses the first four principal components of ancestry to predict an individual's raw polygenic score for breast cancer. We then used residuals from this model to create an ancestry-adjusted

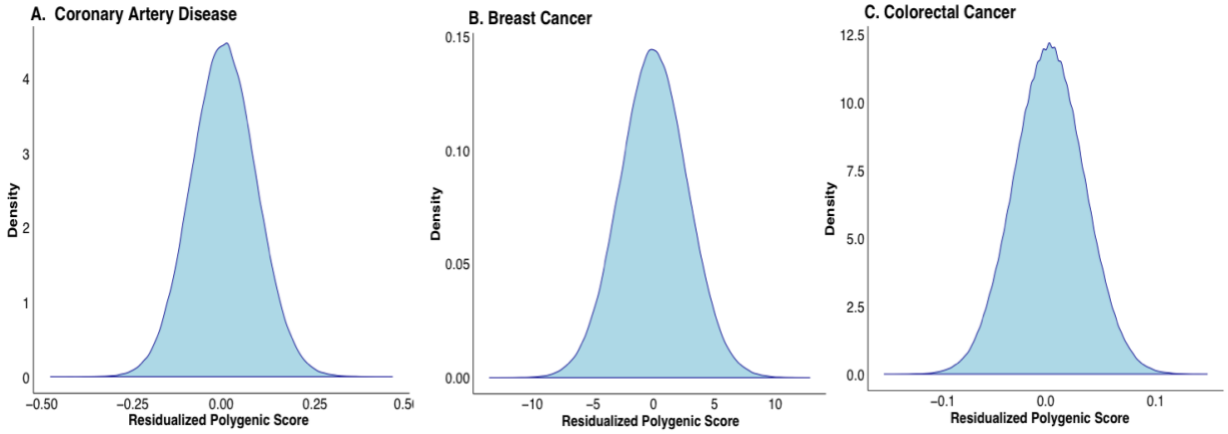
reference distribution for each individual.[10] Percentiles were defined based on the distribution of the residuals of the 19,264 participants.

Assessment of linearity in the polygenic score - disease relationships

To evaluate whether the relationships between polygenic scores and disease risk are linearly associated in the 49,738 participants of the UK Biobank, we used two approaches. First, we used a likelihood-ratio test to assess whether nonlinear terms can better explain the risk model by comparing a logistic regression model with polygenic score residuals as a single linear predictor (Model 1) to a polynomial model with additional higher degree nonlinear terms (square and cube) of polygenic score residuals (Model 2) (Table S5). Second, we checked model goodness-of-fit by visualizing the predicted risk with the observed risk through the Hosmer-Lemeshow method (Figure S2). We plotted the observed to expected probability of disease in 20 groups of polygenic score percentiles (5% each). We then modeled the odds ratio of disease in carriers and non-carriers as a function of polygenic score (total 100 risk groups partitioned by percentiles) conditioned on covariates as age, gender (for coronary artery disease and colorectal cancer only), and the first four principal components of ancestry in a logistic regression model.

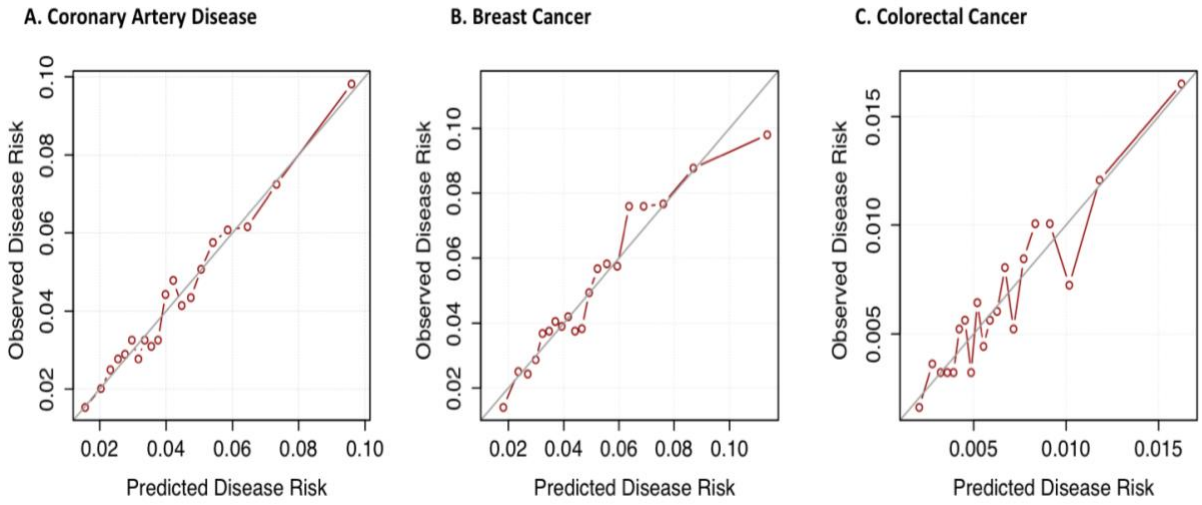
Supplementary figures

Figure S1. Reference polygenic score distributions in the UK Biobank



Shown here are the reference distributions of the polygenic scores used in the study for coronary artery disease in panel A, breast cancer in panel B, and colorectal cancer in panel C in the entire UK biobank population (n=486,477). In this previously described approach to minimize population stratification,[17] we regressed out the top four principal components of ancestry from the raw polygenic risk scores, and made these scores as ancestry-corrected polygenic risk scores (*PS_RES*).

Figure S2. Plot of observed vs. predicted risk of disease using the Hosmer-Lemeshow method



For UK Biobank participants (n=49,738), shown here is model goodness-of-fit evaluation of polygenic background to disease risk by Hosmer-Lemeshow method for coronary artery disease in panel A, breast cancer in panel B, and colorectal cancer in panel C. A single linear predictor of ancestry-corrected polygenic risk score was used to predict the disease risk by a logistic regression model for each disease separately. Each plot shows the observed to predicted probability of disease in 20 groups of polygenic score percentiles (5% each).

Supplementary Tables

Table S1. Baseline characteristics of the 49,738 UK Biobank participants

Characteristic	Total Subjects (N= 49,738)
Age, mean (SD), y	57.1 (8.0)
Female sex, n (%)	27,144 (54.6)
Race, n (%)	
White	46,428 (93.3)
Black	1007 (2)
Asian	1232 (2.5)
Other	1066 (2.2)
Diseases	
Coronary artery disease, n (%)	2120 (4.3)
Breast cancer, n (%)	1369 (2.8)
Colorectal cancer, n (%)	322 (0.65)
Monogenic mutation carriers	
FH variant, n (%)	131 (0.26)
HBOC variant, n (%)*	235 (0.47)
LS variant, n (%)	76 (0.15)

FH= familial hypercholesterolemia ; HBOC= hereditary breast and ovarian cancer syndrome ; LS= Lynch syndrome ; * The number of HBOC mutation carriers in females was 116

Table S2. Prevalent and incident Tier 1 disease among the 49,738 UK Biobank participants

Disease	Total, N (%)	Prevalent, N (%)	Incident, N (%) [s]
Coronary artery disease	3717(7.5)	2120 (4.3)	1597 (3.4)
Colorectal cancer	687 (1.4)	322 (0.65)	365 (0.7)
Breast cancer *	1930 (7.1)	1358 (5.0)	572 (2.2)

[s] The estimate of incident disease excludes prevalent cases at baseline

* Breast cancer is only shown in 27,144 female participants.

Table S3. Likelihood-ratio test for linearity assessment of polygenic score-disease relationships

Disease	Model 1	Model 2	p-value
Coronary artery disease	$CAD \sim PS\_RES_{CAD}$	$CAD \sim PS\_RES_{CAD} + (PS\_RES_{CAD})_2 + (PS\_RES_{CAD})_3$	0.98
Breast cancer	$BrCA \sim PS\_RES_{BrCA}$	$BrCA \sim PS\_RES_{BrCA} + (PS\_RES_{BrCA})_2 + (PS\_RES_{BrCA})_3$	0.07
Colorectal cancer	$CRC \sim PS\_RES_{CRC}$	$CRC \sim PS\_RES_{CRC} + (PS\_RES_{CRC})_2 + (PRS\_RES_{CRC})_3$	0.69

CAD = coronary artery disease, BrCA= breast cancer, CRC = colorectal cancer
PS_RES is the regression residual of regressing disease-specific polygenic score on the first four principal components of ancestry.

Supplementary Appendix References

1. Neben CL, Zimmer AD, Stedden W, et al. Multi-Gene Panel Testing of 23,179 Individuals for Hereditary Cancer Risk Identifies Pathogenic Variant Carriers Missed by Current Genetic Testing Guidelines. J Mol Diagn 2019;21(4):646–57.
2. Hout CVV, Tachmazidou I, Backman JD, et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. bioRxiv 2019;572347.
3. Khera AV, Chaffin M, Wade KH, et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. Cell 2019;177(3):587-596 e9.
4. Fisher S, Barry A, Abreu J, et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol 2011;12(1):R1.
5. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25(14):1754–60.
6. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al 2013;11(1110):11.10.1-11.10.33.
7. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297–303.
8. Lin MF, Rodeh O, Penn J, et al. GLnexus: joint variant calling for large cohort sequencing. bioRxiv 2018;343970.
9. Regier AA, Farjoun Y, Larson DE, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun [Internet] 2018 [cited 2019 Oct 7];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6168605/
10. Karczewski KJ, Francioli LC, Tiao G, et al. Variation across 141,456 human exomes and

genomes reveals the spectrum of loss-of- function intolerance across human protein-coding genes. bioRxiv 2019;

11. Patel, Aniruddh. Penetrance of Monogenic Variants for CDC Tier 1 Genomic Conditions in a Prospective Population Cohort. bioRxiv
12. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 2014;30(20):2843–51.
13. Richards S, Aziz N, Bale S, et al. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med Off J Am Coll Med Genet 2015;17(5):405–24.
14. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 2018;
15. Huyghe JR, Bien SA, Harrison TA, et al. Discovery of common and rare genetic risk variants for colorectal cancer. Nat Genet 2019;51(1):76–87.
16. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet 2019;104(1):21–34.
17. Khera AV, Chaffin M, Zekavat SM, et al. Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. Circulation 2019;139(13):1593–602.
18. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera AV. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. bioRxiv 2019;716977.