

Pathway Polygenic Risk Scores (pPRS) for the Analysis of Gene-environment Interaction

W. James Gauderman¹, Yubo Fu¹, Bryan Queme², Eric Kawaguchi¹, Yinqiao Wang¹, John Morrison¹, Hermann Brenner^{3,4}, Andrew Chan^{5,6}, Stephen B. Gruber⁷, Temitope Keku⁸, Li Li⁹, Victor Moreno^{10,11,12,13}, Andrew J Pellatt¹⁴, Ulrike Peters^{15,16}, N. Jewel Samadder¹⁷, Stephanie L. Schmit^{18,19}, Cornelia M. Ulrich, PhD^{20,21}, Caroline Um²², Anna Wu²³, Juan Pablo Lewinger¹, David A. Drew^{5,6}, Huaiyu Mi²

Affiliations

1. Division of Biostatistics, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA, USA.
2. Division of Bioinformatics, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA, USA.
3. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.
4. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.
5. Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA.
6. Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA.
7. Center for Precision Medicine and Department of Medical Oncology, City of Hope National Medical Center, Duarte, CA, USA.
8. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.
9. Department of Family Medicine, UVA Comprehensive Cancer Center, UVA School of Medicine, Charlottesville, VA, USA.
10. Oncology Data Analytics Program, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, 08908 Barcelona, Spain.
11. Colorectal Cancer Group, ONCOBELL Program, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, 08908 Barcelona, Spain.
12. Department of Clinical Sciences, Faculty of Medicine and Health Sciences and Universitat de Barcelona Institute of Complex Systems (UBICS), University of Barcelona (UB), L'Hospitalet de Llobregat, 08908 Barcelona, Spain.
13. Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), 28029 Madrid, Spain.
14. Intermountain Health, Salt Lake City, UT, USA.

15. Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, Washington, USA.
16. Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, USA.
17. Mayo Clinic Comprehensive Cancer Center, Phoenix, AZ, USA.
18. Genomic Medicine Institute, Cleveland Clinic, Cleveland, OH, USA.
19. Population and Cancer Prevention Program, Case Comprehensive Cancer Center, Cleveland, OH, USA.
20. Huntsman Cancer Institute, Salt Lake City, UT, USA.
21. Department of Population Sciences, University of Utah, USA.
22. Department of Population Science, American Cancer Society, Atlanta, GA, USA.
23. Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA, USA.

Abstract

A polygenic risk score (PRS) is used to quantify the combined disease risk of many genetic variants. For complex human traits there is interest in determining whether the PRS modifies, i.e. interacts with, important environmental (E) risk factors. Detection of a PRS by environment (PRS x E) interaction may provide clues to underlying biology and can be useful in developing targeted prevention strategies for modifiable risk factors. The standard PRS may include a subset of variants that interact with E but a much larger subset of variants that affect disease without regard to E. This latter subset will ‘water down’ the underlying signal in former subset, leading to reduced power to detect PRS x E interaction. We explore the use of pathway-defined PRS (pPRS) scores, using state of the art tools to annotate subsets of variants to genomic pathways. We demonstrate via simulation that testing targeted pPRS x E interaction can yield substantially greater power than testing overall PRS x E interaction. We also analyze a large study (N=78,253) of colorectal cancer (CRC) where E = non-steroidal anti-inflammatory drugs (NSAIDs), a well-established protective exposure. While no evidence of overall PRS x NSAIDs interaction ($p=0.41$) is observed, a significant pPRS x NSAIDs interaction ($p=0.0003$) is identified based on SNPs within the TGF- β / gonadotropin releasing hormone receptor (GRHR) pathway. NSAIDs is protective (OR=0.84) for those at the 5th percentile of the TGF- β /GRHR pPRS (low genetic risk, OR), but significantly more protective (OR=0.70) for those at the 95th percentile (high genetic risk). From a biological perspective, this suggests that NSAIDs may act to reduce CRC risk specifically through genes in these pathways. From a population health perspective, our result suggests that focusing on genes within these pathways may be effective at identifying those for whom NSAIDs-based CRC-prevention efforts may be most effective.

Author Summary

The identification of polygenic risk score (PRS) by environment (PRSxE) interactions may provide clues to underlying biology and facilitate targeted disease prevention strategies. The

standard approach to computing a PRS likely includes many variants that affect disease without regard to E, reducing power to detect PRS x E interactions. We utilize gene annotation tools to develop pathway-based PRS (pPRS) scores and show by simulation studies that testing pPRS x E interaction can yield substantially greater power than testing PRS x E, while also integrating biological knowledge into the analysis. We apply our method to a large study of colorectal cancer to identify a significant pPRS x NSAIDs interaction ($p=0.0003$) based on SNPs within the TGF- β / gonadotropin releasing hormone receptor (GRHR) pathway. Our findings suggest that focusing on genetic susceptibility within biologically informed pathways may be more sensitive for identifying exposures that can be considered as part of a precision prevention approach.

Introduction

Gene-environment (GxE) interactions likely play an important role in the etiology of most complex human traits¹. A GxE analysis aims to identify genetically defined subsets of the population that may be more sensitive to adverse or protective effects of an exposure on disease risk. Alternatively, one can view G x E interaction as investigating whether a particular exposure stimulates or suppresses the effect of a gene on disease risk. The power to detect GxE interactions, particularly in the context of a genomewide scan, is lower than the power to detect similarly-sized genetic or environmental main effects². Identification of actionable GxE interactions is essential to precision medicine approaches that are expected to transform the future of medicine, particularly for primary prevention of diseases.

A polygenic risk score (PRS) is commonly used to summarize the overall effect of a collection of identified genetic variants on a particular trait. The variants used to construct the PRS can be focused on a relatively small set identified by a prior GWAS or a much larger set that captures genome-wide genetic variation. The PRS can be used to characterize the total trait variance attributable to discovered variants or to identify specific subsets of the population likely to be at highest risk for disease³. PRS can also be used in Mendelian randomization analysis when the disease risk factor of interest is itself predictable based on prior GWAS-discovered variants⁴.

Recently, many investigators have utilized PRS x E analysis to study gene-environment interactions for a wide range of traits, including lung cancer⁵, diabetes⁶, ADHD⁷, and cardiovascular disease⁸. Compared to single-variant GxE analysis, PRS x E analysis may provide increased power because it focuses on known disease-related variants and it integrates the signals across those variants into a potentially more informative single measure of genetic susceptibility⁹. Detecting a PRS x E interaction will allow us to answer questions such as: Does the effect of a particular exposure on disease risk vary depending on overall genetic

susceptibility? Do we need to consider specific exposures when making PRS-based risk predictions? Is there a particularly high-risk subgroup, defined by both genetic susceptibility and exposure, for whom targeted prevention (e.g. early screening) may be indicated?

Despite these advantages, a potential difficulty in identifying PRS x E is that standard construction of the PRS includes all GWAS-significant variants or a very large set of genomewide variants. Environmental factors likely work to affect disease risk by altering the functioning or expression of genes within specific pathways. Examples include smoking affecting DNA repair pathways to alter lung cancer risk¹⁰ and red meat affecting inflammatory response pathways to affect colorectal cancer risk¹¹. While a standard PRS may include several variants within an exposure-relevant pathway, its standard construction will tend to ‘water down’ the specific signals most important for identifying the interaction(s).

To overcome this challenge, we propose the use of pathway polygenic risk scores (pPRS) in gene-environment interaction analyses. Relative to a PRS, a pPRS may include a greater proportion of disease-related SNPs that individually or in combination interact with a particular exposure, and which in turn should provide greater power for detecting pPRS x E compared to PRS x E. We will describe the use of available functional annotation databases to define subsets of PRS SNPs according to their known pathway affiliation. Multiple pPRS can be constructed, each corresponding to a particular pathway and utilizing a subset of the overall collection of PRS SNPs. The use of pathway-specific PRS has been described for classifying disease subtypes¹²⁻¹⁴ and enhancing drug target discovery¹⁵, but to our knowledge not for identifying pPRS x E interactions. To illustrate our approach, we analyze PRS x E and pPRS x E interactions in a large study of colorectal cancer, focusing on over 200 GWAS-identified SNPs and a well-established protective exposure, non-steroidal anti-inflammatory drug (NSAID) use.

Results

Simulations

We designed a simulation study to determine whether power to detect pPRS x E interaction may be higher than for PRS x E interaction, and if so, under what conditions one may expect greater power. Briefly, we simulated 1,000 SNPs, of which 20 were assumed to affect disease (D) risk and 980 to have no effect on D. We also simulated a binary exposure (E) and generated 5 of the 20 SNPs to also have a GxE effect on D. We assumed 5 of the 1,000 SNPs fell within a pathway and varied how many of those 5 pathway SNPs overlapped with the 5 GxE SNPs, the 15 other disease-causing SNPs, and the remaining 980 null SNPs. We replicated the simulation 1,000 times and estimated power based on the proportion of replicates in which we detected interaction based on analysis of PRS x E vs. pPRS x E. Additional details of the simulation design, as well as demonstration that Type I error is preserved, are provided in Materials and Methods.

Across a wide range of simulated scenarios, power to detect interaction is greater for pPRSxE than for PRSxE (Table 1). With 20 simulated disease-causing SNPs, there was a cross-replicate average of 18.2 SNPs identified by GWAS and used for constructing the overall PRS, including an average of 4.7 of those 5 SNPs simulated to have a GxE interaction. Power to detect PRSxE interaction using the overall PRS ranged between 41% and 45% across multiple scenarios. When the 5 SNPs simulated to have a GxE effect were synonymous with the 5 SNPs in the pathway, power of the pPRSxE test was substantially higher (90%, scenario 1). This demonstrates the increased efficiency in focusing on a well-chosen subset of SNPs and corresponding pPRSxE test rather than attenuating the interaction signal in an overall PRSxE test.

We also considered simulation scenarios in which only a subset of the 5 pathway SNPs overlapped with the 5 GxE SNPs. These included scenarios in which the pathway SNPs without a GxE effect either did (Table 1, Scenarios 2-5) or did not (Scenarios 6-9) have a main (G only) effect on the trait. When the 5 pathway SNPs include 4 with true GxE and 1 G-only (scenario 2) or 3 GxE and 2 G-only (scenario 3), power of the pPRSxE test was still greater (74%, 47%, respectively) than the PRSxE test. However, with 2 GxE and 3 G-only (scenario 4) or 1 GxE and 4 G-only (scenario 5), power of the pPRSxE was lower (23%, 7%, respectively). By comparison, when the 5 pathway SNPs included 4 with true GxE and 1 with no effect on the trait (scenario 6), power was 84%, larger than the 74% when the non-GxE SNP had a G-only effect (scenario 2). This is because in scenario 6 the non-GxE SNP likely is not discovered in the initial GWAS and thus is not used in forming the pPRS (or PRS) score, and therefore is not attenuating the signal in the remaining GxE SNPs. This trend is further exemplified by the corresponding higher powers in scenarios 7, 8, and 9 compared to scenarios 3, 4, and 5, respectively.

Table 1: Power to detect polygenic risk score by E interactions

Sim	# Pathway SNPs	Pathway-SNP Effects on D			Power		
		GxE - D	G - D only	No effect	PRS x E	pPRS x E	npPRS x E
1	5	5	0	0	44%	90%	2%
2	5	4	1	0	41%	74%	4%
3	5	3	2	0	41%	47%	7%
4	5	2	3	0	45%	23%	17%
5	5	1	4	0	45%	7%	28%
6	5	4	0	1	41%	84%	4%
7	5	3	0	2	41%	69%	7%
8	5	2	0	3	45%	52%	14%
9	5	1	0	4	45%	27%	25%

Simulated power based on 1,000 replicates. Each replicate includes 15 SNPs with a G-only effect on D and 5 SNPs with a GxE effect on D. There are 5 SNPs in the pathway.

Each simulation scenario varies the number of pathway SNPs that overlap with the GxE SNPs (GxE-D), G-only SNPs (G-D), and no-effect SNPs.

Power is the proportion of replicates in which the null hypothesis of no interaction is rejected when the polygenic score is based on all GWAS significant SNPs (PRS x E), GWAS SNPs in the pathway (pPRS x E) or GWAS SNPs not in the pathway (npPRS x E).

Colorectal Cancer (CRC) Application

The most recent and largest GWAS of CRC described a total of 204 previously identified and novel autosomal SNPs that reached genome-wide significance¹⁶. We investigated whether PRS and pPRS formed from these SNPs interact with use of aspirin or non-steroidal anti-inflammatory drugs (NSAIDs) use, a factor well-established to reduce CRC risk¹⁷⁻¹⁹. We used data from the Functionally Informed Gene-environment Interaction (FIGI) study, a consortium of 45 studies that includes 78,253 subjects (33,937 cases, 44,316 controls) with complete data on NSAIDs, genotypes, and covariate data¹⁹. Adjusting for covariates, the NSAIDs main effect on CRC is OR=0.76 (95% C.I. 0.74, 0.79). Although NSAIDs is a protective factor on average, there are risks associated with regular use, such as gastrointestinal bleeding, that necessitate a precision prevention approach. This is one motivation for exploring a precision prevention approach for NSAIDs based on possible modification by genetic susceptibility.

We constructed an overall PRS by first applying logistic regression within the FIGI sample to model CRC as a function of the 204 GWAS SNPs, with adjustment for study, sex, age, and three global ancestry PCs (see Materials and Methods). The SNP-specific log-odds ratios estimated from this model were used as the weights $[w]$ to construct a PRS_i , $i=1, \dots, N$ for each study subject. To construct pPRS, we first used ANNOQ²⁰ which successfully annotated 189 of the 204 SNPs to 265 protein-coding genes. The remaining 15 SNPs were mapped to non-coding genes and are ignored in this analysis. Application of PANTHER²¹ annotated 66 of the 265 genes to a total of 50 pathways (Figure 1), with pathways for the remaining 199 genes not identified. Among the 50 pathways, four of them included more genes than expected by chance alone at a false discovery rate (FDR) of 5%, identified by a Fisher's Exact test in PANTHER (Table 2). These included the TGF- β signaling pathway (raw $p=5.8 \times 10^{-6}$), Alzheimer disease presenilin pathway ($p=5.8 \times 10^{-5}$), Gonadotropin-releasing hormone receptor pathway ($p=4.8 \times 10^{-5}$), and the Cadherin signaling pathway ($p=1.38 \times 10^{-3}$). A total of 30 of the 204 SNPs were

annotated to genes in these pathways. Subsets of the above PRS weights were utilized to construct the corresponding four pPRS scores. Annotated genes in the TGF- β signaling (TGF- β) pathway and Gonadotropin-releasing hormone receptor (GRHR) pathways are highly overlapped (Figure 2A), as are genes in the Cadherin signaling (CADH) and Alzheimer's disease presenilin(ALZ) pathways (Figure 2B). These overlaps lead to significant correlations between the computed pPRS scores for TGF- β and GRHR ($R^2=0.58$) and for CADH and ALZ ($R^2=0.71$). Given this, we also constructed two additional pPRS scores based on SNPs within the combined subsets of TGF- β /GRHR genes and CADH/ALZ genes, respectively.

The estimated GxE odds ratio (OR_{GxE}) for the overall PRS x NSAIDs interaction is 0.99 and is not statistically significant ($p=0.41$, Table 3). We also did not observe significant pPRS x E interactions for the CADH and ALZ pathways. However, the pPRS x NSAIDs interaction was significant for both the TGF- β ($OR_{GxE}=0.96$, $p=0.0069$) and GRHR ($OR_{GxE}=0.96$, $p=0.016$) pathways. The TGF- β and GRHR pathways combined include 20 of the 204 SNPs (Figure 2A). The pPRS x NSAIDs interaction is more pronounced ($OR_{GxE}=0.94$, $p=0.0003$) based on the pPRS formed from this joint set of TGF- β and GRHR SNPs (Table 3). This estimate can be interpreted as an additional 0.94 protective effect of NSAIDs on CRC risk per increase of 1 standard deviation in the combined TGF- β /GRHR pPRS. To further explore this interaction, we used the model to predict the NSAIDs effect on CRC at various percentiles of the TGF- β /GRHR pPRS (Figure 3). For those at the 5th percentile of the pPRS (low risk), the estimated NSAIDs OR is 0.84 (0.79, 0.89) while at the 95th percentile (high risk), it is 0.70 (0.65, 0.74). Put another way, regular NSAIDs use is predicted to reduce CRC risk by 16% for those at low risk based on the TGF- β /GRHR pPRS and by 30% for those at high TGF- β /GRHR pPRS risk.

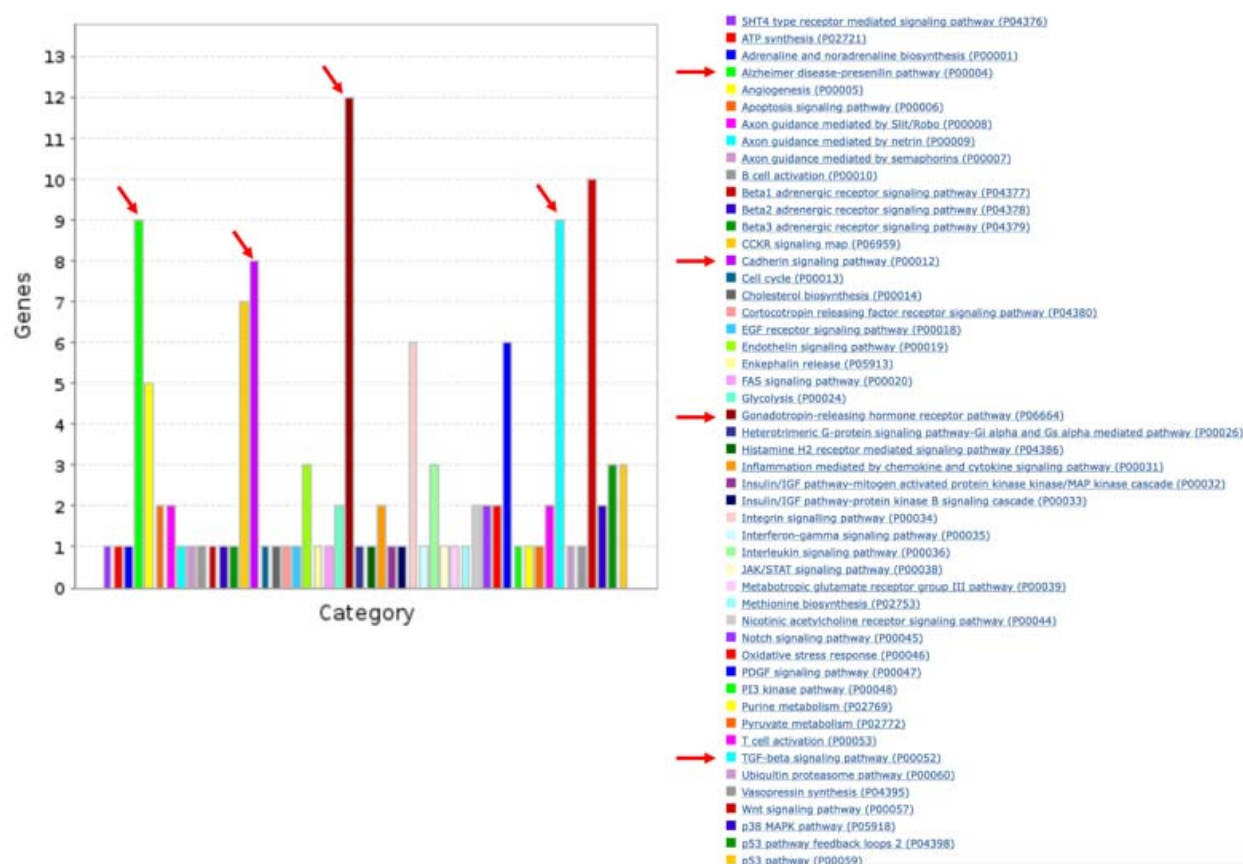


Figure 1: Number of genes and pathways represented among the 204 CRC-associated SNPs. Arrows indicate the four pathways with significant over-representation compared to expected (Table 2).

Table 2: Four pathways significantly over-represented among the 204 CRC-related SNPs

	Homo sapiens (REF)	AnnoQ_uniq_genes					
PANTHER Pathways	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
TGF-beta signaling pathway	100	9	1.29	6.99	+	5.82E-06	4.66E-04
Alzheimer disease-presenilin pathway	127	9	1.64	5.50	+	4.01E-05	2.14E-03
Gonadotropin-releasing hormone receptor pathway	231	12	2.97	4.03	+	4.83E-05	1.93E-03
Cadherin signaling pathway	163	8	2.10	3.81	+	1.27E-03	4.06E-02

Column 2 shows the total number of genes in Homo sapiens annotated to the indicated pathway.

Column 3 shows the number of genes from the pathway that are among the genes linked by application of AnnoQ to the 204-CRC associated SNPs

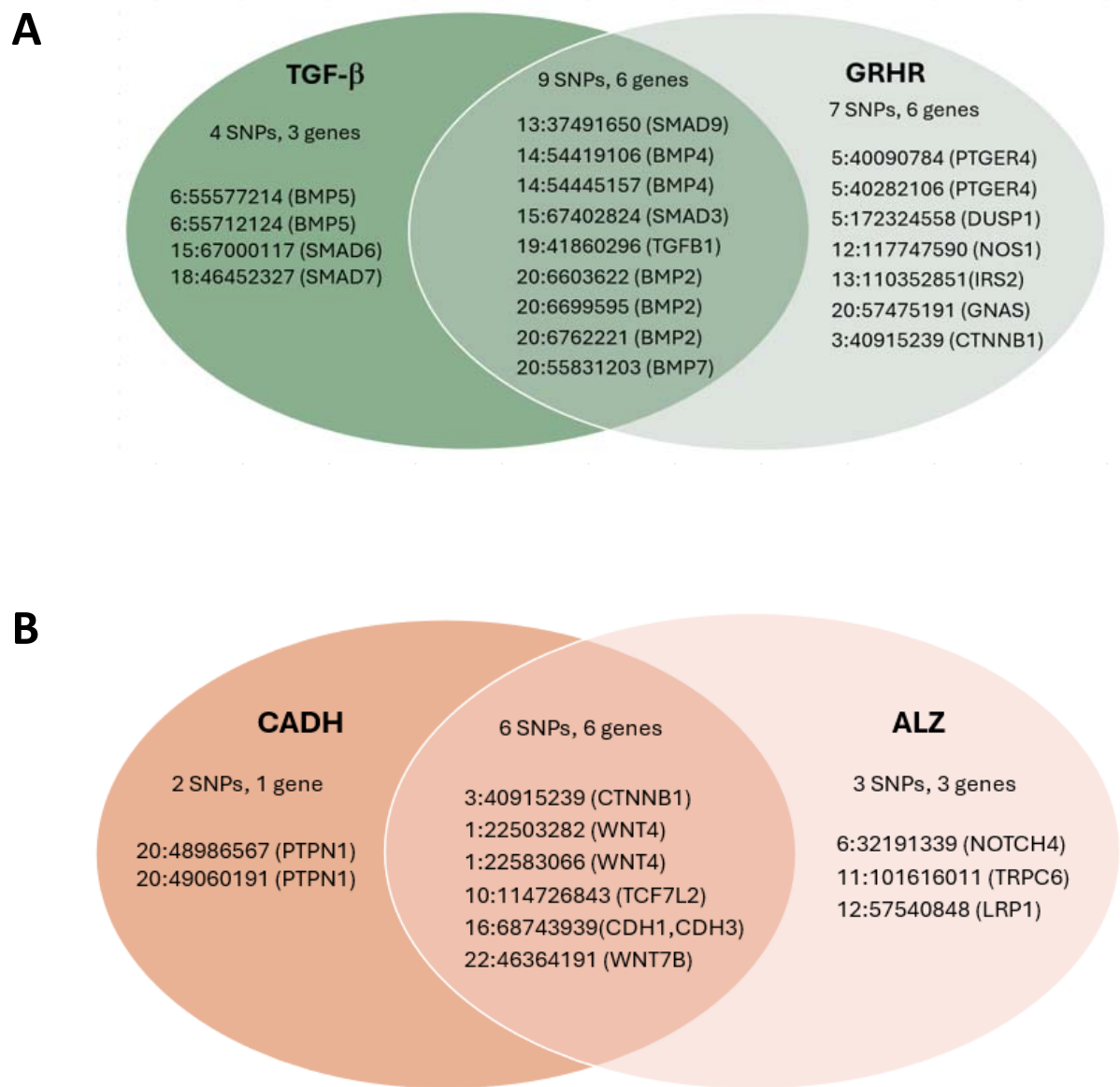


Figure 2: A. Subset of 204 CRC-associated SNPs annotated to genes within the TGF- β and/or the Gonadotropin releasing hormone receptor (GRHR) pathways. **B.** Subset of 204 CRC-associated SNPs annotated to genes within the Cadherin signaling (CADH) and/or the Alzheimer's disease-presenilin (ALZ) pathways.

Table 3: Analysis of polygenic risk score x NSAIDs interaction for Colorectal Cancer

PRS Type	PRS		E (NSAIDs use)		PRS x E		
	OR ^a	(95% CI)	OR	(95% CI)	OR	(95% CI)	p-value ^b
PRS: All SNPs*	1.63	(1.61, 1.66)	0.76	(0.74, 0.79)	0.99	(0.95, 1.02)	0.41
<u>4 Pathways^{&}</u>							
pPRS: TGF- β	1.18	(1.16, 1.20)	0.76	(0.74, 0.79)	0.96	(0.93, 0.99)	0.0069
pPRS: Gonadotropin receptor	1.17	(1.15, 1.19)	0.76	(0.74, 0.79)	0.96	(0.93, 0.99)	0.016
pPRS: Cadherin signaling	1.10	(1.09, 1.12)	0.76	(0.74, 0.79)	1.00	(0.97, 1.04)	0.82
pPRS: Alzheimer's presenillin	1.09	(1.08, 1.11)	0.76	(0.74, 0.79)	0.99	(0.96, 1.02)	0.46
<u>2 combined Pathways^{&}</u>							
pPRS: TGF- β and/or Gonadotropin receptor	1.21	(1.19, 1.23)	0.76	(0.74, 0.79)	0.94	(0.92, 0.97)	0.0003
pPRS: Cadherin and/or Alzheimer's presenillin	1.11	(1.10, 1.13)	0.76	(0.74, 0.79)	1.00	(0.97, 1.03)	0.86
PRS Other [#]	1.55	(1.53, 1.58)	0.76	(0.74, 0.79)	1.01	(0.98, 1.04)	0.63

* PRS formed based on 204 GWAS significant SNPs as reported in Fernandez-Rozadilla et al. (2022)

& pPRS based on subsets of the 204 SNPs within the indicated pathway

PRS based on the subset of 174 of the 204 SNPs that are not within any of the indicated pathways

a Odds ratios (OR) are scaled to a 1 s.d. increase for the indicated PRS and compare users to non-users for NSAIDs. All $p < 10^{-10}$.

b p-value for the test of the null hypothesis of no PRS x E interaction.

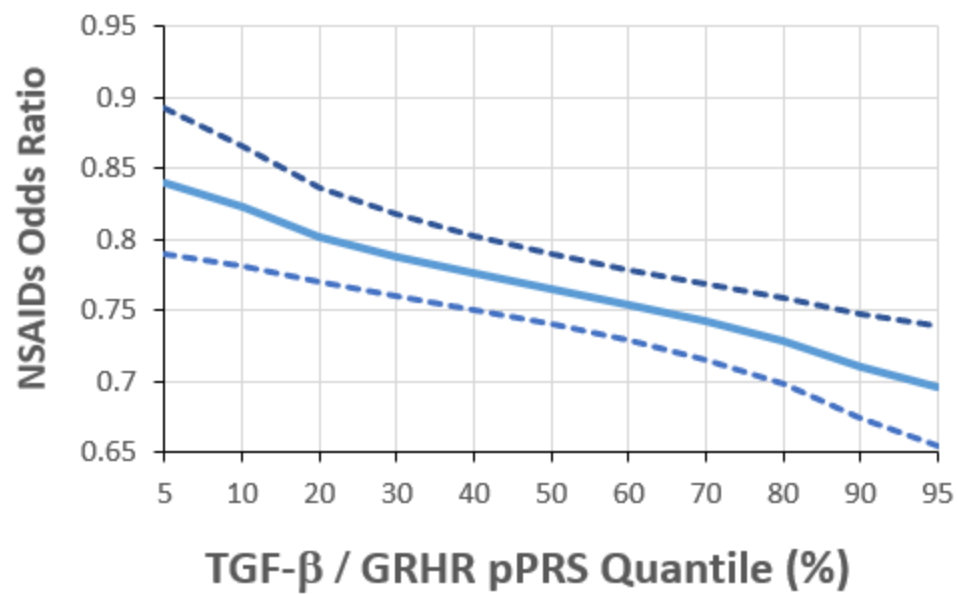


Figure 3: NSAIDs odds ratio for CRC by quantile of the joint TGF-β / GRHR pPRS quantile.

We repeated these analyses utilizing PRS weights obtained from the PGS catalog for the same set of SNPs (PGS-ID 003850). This was done to further evaluate how use of our own data to estimate PRS weights (as above) compared to the more standard approach of using catalog-derived, published weights. Applying the two sets of weights to our analysis sample yielded PRS scores that were very highly correlated for the overall PRS ($R^2=0.9$) as well as for the TGF- β (0.98), GRHR (0.97), CADH (0.97), and ALZ (0.89) pPRS. Not surprisingly, then, results based on PGS catalog weights (Table 4) were very similar to those reported above (Table 3), with similar interaction estimates and levels of significance for TGF- β , GRHR and the joint TGF- β /GRHR pPRS x NSAIDs effects, and non-significant results for the other pathway and overall PRS x NSAIDs tests.

Table 4: Analysis of PGS Catalog derived polygenic risk score x NSAIDs interaction for Colorectal Cancer

PRS Type	PRS		E (NSAIDs use)		PRS x E		
	OR ^a	(95% CI)	OR ^a	(95% CI)	OR	(95% CI)	p-value ^b
PRS: All SNPs*	1.59	(1.56, 1.61)	0.77	(0.74, 0.79)	0.98	(0.95, 1.01)	0.24
<u>4 Pathways^{&}</u>							
pPRS: TGF- β	1.18	(1.16, 1.20)	0.76	(0.74, 0.79)	0.96	(0.93, 0.99)	0.009
pPRS: Gonadotropin receptor	1.17	(1.15, 1.18)	0.76	(0.74, 0.79)	0.96	(0.94, 1.00)	0.021
pPRS: Cadherin signaling	1.10	(1.08, 1.11)	0.76	(0.74, 0.79)	1.00	(0.97, 1.03)	0.84
pPRS: Alzheimer's presenillin	1.08	(1.07, 1.10)	0.76	(0.74, 0.79)	0.99	(0.96, 1.02)	0.64
<u>2 Combined Pathways^{&}</u>							
pPRS: TGF- β and/or Gonadotropin receptor	1.21	(1.19, 1.23)	0.76	(0.74, 0.79)	0.95	(0.92, 0.98)	0.0004
pPRS: Cadherin and/or Alzheimer's presenillin	1.10	(1.09, 1.12)	0.76	(0.74, 0.79)	1.00	(0.97, 1.03)	0.998
PRS Other [#]	1.51	(1.49, 1.53)	0.77	(0.74, 0.79)	1.00	(0.97, 1.03)	0.957

* PRS formed based on 204 GWAS significant SNPs as reported in Fernandez-Rozadilla et al. (2022)

& pPRS based on subsets of the 204 SNPs within the indicated pathway

PRS based on the subset of 174 of the 204 SNPs that are not within any of the indicated pathways

a Odds ratios (OR) are scaled to a 1 s.d. increase for the indicated PRS and compare users to non-users for NSAIDs. All $p < 10^{-10}$.

b p-value for the test of the null hypothesis of no PRS x E interaction.

Discussion

We have demonstrated by simulation and application to data that forming a PRS based only on a subset of GWAS significant SNPs, specifically a subset defined a priori based on pathway information, has the potential to better identify novel PRS x E interactions. We also demonstrate that power may be reduced using the standard practice of testing PRS x E interaction based only on an overall PRS. This power reduction is likely due to ‘watering down’ the interaction signal with the inclusion of most of the SNPs in the PRS construction that do not have any role in modifying the effect of E on disease. By contrast, the use of external pathway information to form a pPRS has the potential to improve power by focusing on genetic variation within a particular pathway that modifies the E effect. Examination of E effects across quantiles of the pPRS can identify those genetically-defined subsets that are most affected, or protected, by exposure. For example, our analysis of CRC suggests that although NSAIDs use is generally beneficial for all, those with the highest TGF- β /GRHR pathway PRS experience a significantly greater reduction in CRC relative risk with regular NSAIDs use. This result both adds to the overall preventive evidence for NSAIDs on CRC risk and suggests possible biological pathways that are involved in this action.

The use of pPRS in interaction testing relies on external information to identify the pathways corresponding to a particular set of SNPs. In this paper, we focused on the set of GWAS significant SNPs, but one could relax the criteria to include SNPs that have GWAS p-value that achieves a lesser threshold (e.g. 5×10^{-5}) or to a wider collection without regard to GWAS significance. As has been previously shown, a GxE interaction typically induces a direct disease-gene (DG) association²²⁻²⁵, and so requiring some level of DG association to be included in PRSxE or pPRSxE analysis is reasonable. In our application to CRC, we created a workflow that utilized AnnoQ to annotate SNPs to genes and PANTHER to annotate genes to pathways. We recognize that there are alternative tools/databases that could be employed, for

example Reactome (*reactome.org*) or Gene Ontology (*geneontology.org*), and that different workflows would likely result in pathway assignments that do not fully overlap. A particular application of pPRS x E analysis could consider the use of multiple workflows, each using different tools/databases, to evaluate the sensitivity of findings to specific pathway definitions and corresponding SNP/gene assignments.

An ancillary finding in this paper is the demonstration that one can construct a PRS or pPRS in three different ways if the ultimate focus is a valid test of interaction. Approach #1 (Methods), i.e. to obtain existing PRS weights from the PGS catalog, is the one most often used. This has the advantages that the weights are typically estimated using a large and independent dataset, and that one can apply the weights to their data to estimate both PRS main and interactive effects. A potential disadvantage, however, is that the data used to generate the PGS weights may come from a population(s) that do not represent the sample used for PRS x E analysis. It is well known that cross-population application of PRS for main effects can lead to poor estimation, and the same will hold for analysis of PRS x E interactions. The advantage of Approach #2 (Methods) is that it leverages the discovery of SNPs in a larger, independent population, but tailors the weights used in PRS construction to the specific population being studied for interaction. Of course, this is also not free of cross-population issues if the discovered SNPs in the independent population are not representative of the SNPs/genes affecting the trait in the study population. Approach #3, in which the study sample is used to discover SNPs and estimate weights, is perhaps the cleanest from the standpoint of population heterogeneity but may suffer from reduced power to discover SNPs relative to larger independent studies. As we demonstrated in our CRC analysis, the flexibility to use any of these approaches for valid interaction testing provides the opportunity to evaluate the robustness of PRSxE and/or pPRSxE findings to the choice of PRS SNPs and weights.

Our results highlight that pPRSxE can identify pathways with functional relevance to the exposure's putative mechanisms of action. In this case, we provide evidence that the protective effect of NSAIDs on CRC risk is modified by variation in the TGF- β and GRHR pathways. While aspirin and other NSAIDs primary inhibitory activity on PTGS1/2 (or COX1/2) have long been hypothesized as central mechanisms of for their anticancer effects, the overall mode of action is still not yet clear. Several lines of functional evidence have supported a role for the TGF- β superfamily in mediating aspirin/NSAIDs protective effects against CRC²⁶, particularly in models of mismatch repair deficient CRC²⁷. Long-term follow-up of the CAPP2 randomized, placebo-controlled trial conclusively demonstrated that aspirin is protective against CRC among patients with Lynch syndrome²⁸, also known as hereditary non-polyposis colon cancer resulting from pathogenic variants within DNA mismatch repair genes, suggesting that NSAID protection may also extend to those with sporadic mismatch repair deficient tumors. TGF- β has also been demonstrated to induce *HPGD*²⁷, a prostaglandin-degrading enzyme with tumor suppressor activity that works as a catabolic antagonist for PTGS-2 activity²⁹. Moreover, HPGD mucosal gene expression has been demonstrated to stratify individuals that may be more likely to experience a preventive benefit from aspirin use³⁰. While other TGF- β superfamily members like GDF15 have been proposed as potential markers for precision prevention of CRC with NSAIDs¹⁸, the role for bone morphogenetic proteins (BMPs) and SMAD family proteins in NSAID chemoprotection are less well established than they are for other agents, like metformin³¹, or other physiologic processes, like osteogenic differentiation^{32,33}. Similarly, functional evidence is limited for a specific role of Gonadotropin-receptor pathway overall in NSAIDs mechanisms of action. However, of those genes included in the pPRS score, prior evidence links NSAID anti-cancer activity with β -catenin (CTNNB1³⁴⁻³⁷), GNAS³⁸, and PTGER4¹⁹, the extracellular receptor for PGE₂, the major downstream prostanoid produced by PTGS-2. Combined, these results highlight that a pPRSxE approach may identify additional network nodes with potential functional relevance for future mechanistic interrogation.

We have shown that leveraging prior GWAS results combined with biological information to construct subsets of SNPs in pPRS x E tests has the potential to improve power compared to SNPxE or overall PRSxE tests. An additional advantage of the pPRS x E analysis is that it may strengthen the evidence for a potential biological mechanism, via the involved pathway, by which E affects the outcome. Although we have focused on SNP subsets based on pathway information, we recognize there are other sources of information that could be used to create subsets. For example, subsets could be formed based on information on SNP-expression in a relevant tissue or cell type, or based on SNP associations with traits related to the trait of interest. Future research is needed to examine the robustness of pPRSxE analyses to the choice of annotation workflow, to the approach to creating subsets, and to demonstrate whether pPRS can be used to successfully identify novel gene-environment interactions for other complex traits.

Materials and Methods

Notation and Standard GxE Analysis

Let D_i denote a disease indicator for subject i , $i=1, \dots, N$, E_i an exposure of interest, and $\mathbf{Z}_i$ a vector of adjustment covariates (e.g. age, sex, ancestry principal components). Assume one or more GWAS has been conducted, yielding a set $\mathbf{G}=[G_1, G_2, \dots, G_M]$ of trait associated SNPs, for example those with $p < 5 \times 10^{-8}$ for the test of SNP vs. D association. Assume further that a case-control sample has been obtained, with complete data for $D, E, \mathbf{Z}$, and $\mathbf{G}$ on each subject. For analysis of $G \times E$ interaction with a single SNP, we assume logistic regression model of the form:

$$\text{logit}[\Pr(D|G, E, \mathbf{Z})] = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} G \times E + \boldsymbol{\beta}_z \mathbf{Z} \quad (1)$$

Here β_g denotes the genetic ‘main’ effect quantifying the association between G and D when $E = 0$, β_e is the corresponding environmental main effect, and β_{ge} parameterizes the $G \times E$ interaction effect of primary interest. G is typically coded as the number of minor alleles, 0, 1, or 2 if it is measured or the corresponding expected number if imputed. In practice, we often center both G and E on their respective sample means yielding

$$\text{logit}[\Pr(D|G, E, \mathbf{Z})] = \bar{\beta}_0 + \bar{\beta}_g (G - \bar{G}) + \bar{\beta}_e (E - \bar{E}) + \bar{\beta}_{ge} (G - \bar{G}) \times (E - \bar{E}) + \boldsymbol{\beta}_z \mathbf{Z} \quad (2)$$

Here $\bar{\beta}_g$ parameterizes the G to D association at the mean of E and similarly for $\bar{\beta}_e$. An advantage of this centering is that $\bar{\beta}_g$ and $\bar{\beta}_e$ approximate the ‘marginal’ effects of G and E , for example the direct effect of G on D (γ_g) that is obtained in a GWAS using the model:

$$\text{logit}[\Pr(D|G, \mathbf{Z})] = \gamma_0 + \gamma_g G + \boldsymbol{\gamma}_z \mathbf{Z} \quad (3)$$

The test of $G \times E$ interaction evaluates the null hypothesis $H_0: \beta_{ge} = 0$ and can be based on a Wald, Score, or likelihood-ratio test from either model 1 or model 2, with proper adjustment to the significance level to achieve the desired family-wise error rate. If each of S SNP is tested, the significance level for each SNP is α/S , i.e. based on a Bonferroni correction for S tests subject to overall significance level α .

Formation of the PRS

For a collection of M SNPs, e.g. those previously identified as GWAS significant, the following logistic model is used to estimate the SNP effects in the context of a single joint model:

$$\text{logit}[\Pr(D|G, \mathbf{Z})] = \alpha_0 + \sum_{k=1}^M \alpha_k G_k \quad (4)$$

We define weights $[w_k]$ to be the estimates $[\hat{\alpha}]$ from Model 4. The equation for generating a PRS for the i^{th} individual is

$$\text{PRS}_i = \sum_{k=1}^M w_k G_{ik} \quad (5)$$

Replacing G in Equation 1 by the PRS yields the following model which can be used to estimate and test for PRS $\times$ E interaction:

$$\text{logit}[\Pr(D|\text{PRS}, E, \mathbf{Z})] = \beta_0 + \beta_g \text{PRS} + \beta_e E + \beta_{ge} \text{PRS} \times E + \boldsymbol{\beta}_z \mathbf{Z} \quad (6)$$

PRS Weights

The PRS weights are typically derived from a separate resource. For example, the PGS catalog³⁹ provides weights for over 650 traits, including multiple sets of weights for many of the

traits. It is important that the weights come from independent data resources if the PRS will be used to examine direct risk effects on the disease of interest in the N subjects under study. In other words, if the weights are generated based on the N subjects under study, applying the resulting PRS to the same subjects will result in biased inference of the PRS effect on disease risk. However, we will demonstrate that the same dataset can be used to generate the PRS weights if the focus is on PRS x E interaction. The ability to ‘double use’ the same data to generate and apply the weights relies on the independence between the marginal genetic effects (estimated via Model 3) and the interaction effects (estimated via Model 2). This independence has been shown for tests of single SNPs⁴⁰ and is the basis for several 2-step genomewide GxE scan methods that screen on marginal G effects in Step 1 and use the information to prioritize SNPs for GxE testing in Step 2^{22,24,25,41}. We provide simulations in this paper demonstrating that this independence holds for use of the weights $[w_k]$ derived from Eq. 4 for downstream PRS x E interaction analysis.

Given this independence, there are three Approaches one might consider for generating the $[w_k]$ and corresponding PRS:

1. Obtain $[w_k]$ from prior studies based on one or more independent datasets. As noted above, these could come from the PGS catalog or a specific previous GWAS of the trait of interest. This will provide weights that can be applied to the N subjects under study for use in estimating PRS main and PRS x E interaction effects on D. One must be prepared to assume, however, that the weights generated from the previous population(s) are applicable to the current study population, which may not be reasonable if there are differences in ancestry⁴².

2. Obtain M SNPs from prior GWAS but estimate $[w_k]$ in the current sample that will be used for PRS x E analysis. Again the list of previously identified SNPs could come from the PGS catalog or a specific prior GWAS, but rather than use existing weights, model 5 is applied to the M SNPs in the current data to generate $[w_k]$. The corresponding PRSi, $i=1, \dots, N$, would not provide valid estimates of the PRS main effect but are valid for estimating and testing PRS x E effects. An advantage of this approach is that the weights are computed based on the demographic (e.g. sex, age, ancestry) composition of the current study. The discovery of the set of M SNPs, however, may have been based on different populations with different exposure histories and thus may not fully represent the genetic and GxE contributions in the current sample.

3. Conduct a GWAS on the current sample to both identify M SNPs and compute corresponding $[w_k]$. Compared to approaches 1 and 2, this has the advantage that both the selection of M SNPs and calculation of weights reflect the population structure and exposure characteristics of the current sample. On the other hand, the current sample may be smaller than prior studies and thus have less power to identify important SNPs in the GWAS discovery step.

We will demonstrate the third approach in our simulation and the first two approaches in our application to colorectal cancer.

Pathway PRS

Human genes and their products typically function together within biological pathways to maintain proper cellular functions. SNPs located within or near gene regions have the potential to influence the pathways in which these genes are involved. We assume that the collection of M SNPs used to form the PRS include subsets of SNPs falling within different biological

pathways. To assign each SNP to a pathway, we first use the Annotation Query (AnnoQ) platform²⁰ to derive annotations to Ensembl⁴³ and RefSeq⁴⁴ genes using inferences from ANNOVAR⁴⁵, SnpEff⁴⁶ and VEP⁴⁷. SNPs residing in enhancer regions were linked to their target genes via PEREGRINE⁴⁸. The resulting genes were annotated to pathways using the PANTHER²¹ Classification System (v.18.0)⁴⁹. The set of genes falling within the same pathway were tested for overrepresentation relative to the PANTHER Pathway annotation sets⁵⁰. Each pathway that is significantly over-represented is the focus of pPRS computation and pPRS x E interaction testing.

Assuming that K pathways are identified by the above approach, we define pPRS₁, pPRS₂, ..., pPRS_K to be PRS including only those SNPs within the corresponding pathway. We also let pPRS₀ denote the PRS that includes the subset of M SNPs not annotated to any of the K pathways. Let S_k, k=0,...,K denote the subset of M SNPs included in the kth subset. The pPRS for pathway k is then defined as:

$$pPRS_k = \sum_{j \in S_k} w_j G_j \quad (7)$$

where weights are obtained by one of the three approaches described above. Note that this approach to computing pPRS implicitly assumes that the weights are generated from the full model of D that includes all M SNPs, which has the advantage that the weights are mutually adjusted for one another. To investigate a particular pPRS, Equation 6 can be modified to:

$$\text{logit}[\Pr(D|pPRS_k, E, \mathbf{Z})] = \beta_0 + \beta_g pPRS_k + \beta_e E + \beta_{ge} pPRS_k \times E + \boldsymbol{\beta}_z \mathbf{Z} \quad (8)$$

Alternatively, one can also use a model that includes all pPRS, with form:

$$\text{logit}[\Pr(D|pPRS_k, E, \mathbf{Z})] = \beta_0 + \beta_e E + \boldsymbol{\beta}_z \mathbf{Z} + \sum_{k=0}^K (\beta_{g_k} pPRS_k + \beta_{ge_k} pPRS_k \times E) \quad (9)$$

Additional interactions between pPRS and $\mathbf{Z}$ and/or between E and $\mathbf{Z}$ can also be included to account for potential confounding at the level of the pPRS x E effects⁵¹. We note that it is possible for a particular SNP to be annotated to two or more pathways. In this situation, there will be correlation between two pPRS that include the same SNP(s), which will require care in interpreting the resulting effect estimates.

Simulation Studies

We conducted simulation studies to: 1) evaluate the claim that the same dataset can be used to estimate the PRS weights [w], construct a PRS, and obtain valid estimates and tests of PRS x E interaction, and 2) to compare the power of pPRS x E to PRS x E analysis.

We generate a dataset that includes 5,000 cases and 5,000 controls, with a binary exposure E and 1,000 randomly and independently generated SNPs per subject. We designate $Q=20$ of the SNPs to affect disease risk, with Q_G having only a main G to D effect and $Q_{G \times E}$ having both a main and GxE effect. We further assume that $Q_P = 5$ of the 1,000 SNPs fall within a particular pathway and that Q_{PG} of the pathway SNPs have only main effect and $Q_{PG \times E}$ have a GxE effect. We vary Q_{PG} and $Q_{PG \times E}$ across simulation scenarios. For each simulation scenario, we generate 1,000 replicate datasets and use these to evaluate Type I error and power. We generate each G as a binary variable with 35% population prevalence and E as binary with population prevalence 50%. Conditional on simulated G and E , disease status for each subject was generated according to a random Bernoulli distribution with probability of disease (P_D) given by:

$$P_D = \text{expit}(\delta_0 + \delta_E E + \sum_{k \in Q_G} \delta_{G_k} G_k + \sum_{k \in Q_{G \times E}} \delta_{G \times E_k} G_k \times E) \quad (10)$$

The values of $[\delta_{G_k}]$ were determined using Quanto⁵² to achieve an expected power of at least 90% to detect each of the Q SNPs in a GWAS with adjustment for 1,000 tests. The $[\delta_{G \times E_k}]$ values were set to achieve approximately 10% power to detect GxE interaction for each of the $Q_{G \times E}$ SNPs, assuming 20 SNPs are evaluated for SNP x E interaction post-GWAS.

For each simulated dataset, we conducted a GWAS of the 1,000 SNPs to identify the M that were significant at the $0.05/1,000 = 5 \times 10^{-5}$ level. These M SNPs were used in a model of the form in Equation 4 to generate weights $[w_k]$. We computed the standard PRS based on these M weights using Equation 5, the pathway PRS (pPRS) based on Equation 7 for the subset of M within Q_P , and the non-pathway PRS (npPRS) based on Equation 7 for the subset of M not within Q_P . Each simulation scenario was replicated 1,000 times and we tallied the proportion of replicates in which the null hypothesis of no interaction was rejected for likelihood ratio tests of PRSxE, pPRSxE, and npPRSxE based on Equation 8. This proportion estimated Type 1 error in simulations with $Q_{G \times E}=0$ and power when $Q_{G \times E} > 0$.

Our first set of simulations shows that use of the same data set to run a GWAS, generate PRS weights, and test PRS x E interaction (approach #3, see above) preserves the desired Type I error rate for the interaction test (Table S1). We simulate 20 disease-causing SNPs ($\delta_{G_k} \neq 0$ for $k \in Q_G$) and set $\delta_{G_k \times E} = 0$, for all k (Eq. 11). We tested five methods to identify the SNPs to generate PRS weights: 1) Identify the M SNPs that were significant at the $0.05/1,000 = 5 \times 10^{-5}$ level; 2) identify the M that were significant at the $0.05/10 = 5 \times 10^{-3}$ level; 3) identify the M that were significant at the 0.05 level; 4) include the 20 disease-causing SNPs; and 5) randomly select 10 of the 20 disease-causing SNPs and 10 from the 980 null SNPs. Across all these scenarios, the estimated Type I error rate was within simulation variability of the desired 0.05

level. Since approaches #1 and #2 for generating PRS (see above) are subsets of approach #3, we conclude that their corresponding Type I error rates for PRSxE testing are also preserved.

Table S1: Estimated type I error and interaction effect for PRSxE using different PRS construction methods

PRS SNPs	Type I Error* PRS x E	Estimated $\beta_{PRS \times E}$
GWAS significant SNPs under threshold:		
5×10^{-5}	0.054	-0.0036
5×10^{-3}	0.038	-0.0049
5×10^{-1}	0.056	-0.0033
20 main effects SNPs	0.050	-0.0029
10 main effects SNPs and 10 no-effects SNPs	0.050	0.0033

* Estimated Type I error and interaction effect based on 1,000 simulated replicates. Each replicate includes 20 SNPs with only main effects and no interaction

Data Application: Colorectal Cancer

We compare the above approaches in an analysis of GxE interactions for colorectal cancer (CRC). We use data from the Functionally Informed Gene-environment Interaction (FIGI) study, a consortium that includes 108,649 subjects (51,350 CRC cases and 57,299 controls) drawn from 45 contributing studies. We focus on E=regular use of aspirin/NSAIDs (denoted NSAIDs from hereon), an exposure that has been repeatedly shown to reduce the risk of CRC¹⁷⁻¹⁹. A total of 78,253 subjects (33,937 cases, 44,316 controls) have complete data on NSAIDs use and are included in the analyses. Additional details of the study sample and definition of exposure are provided in Drew et al.¹⁹

The most recent and largest GWAS of CRC identified 204 SNPs that reached genomewide significance¹⁶. We apply the approaches described above to assess evidence that the PRS constructed from these SNPs interacts with NSAIDs to affect CRC risk. The overall PRS was constructed by first applying logistic regression within the FIGI sample to the 204 GWAS SNPs, with adjustment for study, sex, age, and three ancestry PCs (approach #2 described above). The log-odds ratios (“betas”) estimated from this model were used as the weights [w] to construct a PRS_i , $i=1, \dots, N$ for each study subject. To construct pPRS, we first used AnnoQ which successfully annotated 189 of the 204 SNPs to 265 protein-coding genes. The remaining 15 SNPs were mapped to non-coding genes and are ignored in this analysis. Application of PANTHER annotated 66 of the 265 genes to a total of 50 pathways (Figure 1), with pathways for the remaining 199 genes not identified. Among the 50 pathways, four of them included more genes than expected by chance alone at a false discovery rate (FDR) of 5%, identified by a Fisher’s Exact test in PANTHER (Table 2). These included the TGF- β / signaling pathway (raw $p=5.8 \times 10^{-6}$), Alzheimer disease presenilin pathway ($p=5.8 \times 10^{-5}$), Gonadotropin-releasing hormone receptor pathway ($p=4.8 \times 10^{-5}$), and Cadherin signaling pathway ($p=1.38 \times 10^{-3}$). A total of 30 of the 204 SNPs were annotated to genes in these pathways. Subsets of the above PRS

weights were utilized to construct the corresponding four pPRS scores. Logistic regression was used to estimate and test interactions, with adjustment for study, sex, age, and three principal components of ancestry. For each of the four pPRS x E tests, we report p-values unadjusted for multiple comparisons, with the rationale that each pathway-based PRS was constructed in advance using auxiliary information. The subsets of genes in the TGF- β signaling (TGF- β) pathway and Gonadotropin-releasing hormone receptor (GRHR) pathways are highly overlapped (Figure 2), as are genes in the Cadherin signaling (CADH) and Alzheimer's disease presenilin(ALZ) pathways (Figure 3). These overlaps lead to significant correlations between the computed pPRS scores for TGF- β and GRHR ($R^2=0.58$) and for CADH and ALZ ($R^2=0.71$). Given these overlaps, we also constructed two additional pPRS scores based on SNPs within the combined subsets of TGF- β /GRHR genes and CADH/ALZ genes, respectively, and used analogous logistic regression models to test the corresponding pPRS x NSAIDs interactions.

Funding source Acknowledgments

Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (R01 CA059045, R01 CA201407, R01 CA273198). Genotyping/Sequencing services were provided by the Center for Inherited Disease Research (CIDR) contract number HHSN268201700006I and HHSN268201200008I. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA015704. Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant S10OD028685.

The ATBC Study is supported by the Intramural Research Program of the U.S. National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

CLUE II funding was from the National Cancer Institute (U01 CA086308, Early Detection Research Network; P30 CA006973), National Institute on Aging (U01 AG018033), and the American Institute for Cancer Research. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government. Maryland Cancer Registry (MCR) Cancer data was provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health, with funding from the State of Maryland and the Maryland Cigarette Restitution Fund. The collection and availability of cancer registry data is also supported by the Cooperative Agreement NU58DP006333, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

ColoCare: This work was supported by the National Institutes of Health (grant numbers R01 CA189184 (Li/Ulrich), U01 CA206110 (Ulrich/Li/Siegel/Figueiredo/Colditz, 2P30CA015704- 40 (Gilliland), R01 CA207371 (Ulrich/Li)), the Matthias Lackas-Foundation, the German Consortium for Translational Cancer Research, and the EU TRANSCAN initiative.

The Colon Cancer Family Registry (CCFR, www.coloncfr.org) is supported in part by funding from the National Cancer Institute (NCI), National Institutes of Health (NIH) (award U01 CA167551). Support for case ascertainment was provided in part from the Surveillance, Epidemiology, and End Results (SEER) Program and the following U.S. state cancer registries: AZ, CO, MN, NC, NH; and by the Victoria Cancer Registry (Australia) and Ontario Cancer Registry (Canada). The CCFR Set-1 (Illumina 1M/1M-Duo) and Set-2 (Illumina Omni1-Quad) scans were supported by NIH awards U01 CA122839 and R01 CA143237 (to GC). The CCFR Set-3 (Affymetrix Axiom CORECT Set array) was supported by NIH award U19 CA148107 and R01 CA81488 (to SBG). The CCFR Set-4 (Illumina OncoArray 600K SNP array) was supported by NIH award U19 CA148107 (to SBG) and by the Center for Inherited Disease Research (CIDR), which is funded by the NIH to the Johns Hopkins University, contract number HHSN268201200008I. Additional funding for the OFCCR/ARCTIC was through award GL201-043 from the Ontario Research Fund (to BWZ), award 112746 from the Canadian Institutes of Health Research (to TJH), through a Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society (to SG), and through generous support from the Ontario Ministry of Research and Innovation. The SFCCR Illumina HumanCytoSNP array was supported in part through NCI/NIH awards U01/U24 CA074794 and R01 CA076366 (to PAN). The content of this manuscript does not necessarily reflect the views or policies of the NCI, NIH or any of the collaborating centers in the Colon Cancer Family Registry (CCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government, any cancer registry, or the CCFR.

COLO2&3: National Institutes of Health (R01 CA060987).

CPS-II: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II (CPS-II) cohort. The study protocol was approved by the institutional review boards of Emory University, and those of participating registries as required.

CRCGEN: Colorectal Cancer Genetics & Genomics, Spanish study was supported by Instituto de Salud Carlos III, co-funded by FEDER funds –a way to build Europe– (grants PI14-613 and PI09-1286), Agency for Management of University and Research Grants (AGAUR) of the Catalan Government (grant 2017SGR723), Junta de Castilla y León (grant LE22A10-2), the Spanish Association Against Cancer (AECC) Scientific Foundation grant GCTRA18022MORE and the Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), action Genrisk. Sample collection of this work was supported by the Xarxa de Bancs de Tumors de Catalunya sponsored by Pla Director d'Oncologia de Catalunya (XBTC), Plataforma Biobancos PT13/0010/0013 and ICOBIOBANC, sponsored by the Catalan Institute of Oncology. We thank CERCA Programme, Generalitat de Catalunya for institutional support.

DACHS: This work was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, HO 5117/2-1, HE 5998/2-1, KL 2354/3-1, RO 2270/8-1 and BR 1704/17-1), the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany, and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A and 01ER1505B).

DALS: National Institutes of Health (R01 CA048998 to M. L. Slattery).

EDRN: This work is funded and supported by the NCI, EDRN Grant (U01-CA152753).

Harvard cohorts: HPFS is supported by the National Institutes of Health (P01 CA055075, UM1 CA167552, U01 CA167552, R01 CA137178, R01 CA151993, and R35 CA197735), NHS by the National Institutes of Health (P01 CA087969, UM1 CA186107, R01 CA137178, R01 CA151993, and R35 CA197735), and PHS by the National Institutes of Health (R01 CA042182).

Hawaii Adenoma Study: NCI grants R01 CA072520.

Kentucky: This work was supported by the following grant support: Clinical Investigator Award from Damon Runyon Cancer Research Foundation (CI-8); NCI R01CA136726.

LCCS: The Leeds Colorectal Cancer Study was funded by the Food Standards Agency and Cancer Research UK Programme Award (C588/A19167).

MCCS: Melbourne Collaborative Cohort Study (MCCS) cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the Australian Cancer Database.

MEC: National Institutes of Health (R37 CA054281, P01 CA033619, and R01 CA063464).

MECC: This work was supported by the National Institutes of Health, U.S. Department of Health and Human Services (R01 CA081488, R01 CA197350, U19 CA148107, R01 CA242218, and a generous gift from Daniel and Maryann Fong.

NCCCS I & II: We acknowledge funding support for this project from the National Institutes of Health, R01 CA066635 and P30 DK034987.

Study-specific Acknowledgements

CCFR: The Colon CFR graciously thanks the generous contributions of their study participants, dedication of study staff, and the financial support from the U.S. National Cancer Institute, without which this important registry would not exist. The authors would like to thank the study participants and staff of the Seattle Colon Cancer Family Registry and the Hormones and Colon Cancer study (CORE Studies).

CLUE II: We thank the participants of Clue II and appreciate the continued efforts of the staff at the Johns Hopkins George W. Comstock Center for Public Health Research and Prevention in the conduct of the Clue II Cohort Study. Cancer data was provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health, with funding from the State of Maryland and the Maryland Cigarette Restitution Fund. The collection and availability of cancer registry data is also supported by the Cooperative Agreement NU58DP006333, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

CPS-II: The authors express sincere appreciation to all Cancer Prevention Study-II participants, and to each member of the study and biospecimen management group. The authors would like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention's National Program of Cancer Registries and cancer registries supported by the National Cancer Institute's Surveillance Epidemiology and End Results Program. The authors assume full responsibility for all analyses and interpretation of results. The views expressed here are those of the authors and do not necessarily represent the American Cancer Society or the American Cancer Society – Cancer Action Network.

DACHS: We thank all participants and cooperating clinicians, and everyone who provided excellent technical assistance.

EDRN: We acknowledge all contributors to the development of the resource at the University of Pittsburgh School of Medicine, Division of Gastroenterology, Hepatology and Nutrition, Department of Pathology, and Biomedical Informatics.

Harvard cohorts: The study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required. We acknowledge Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital as home of the NHS. The authors would like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention's National Program of Cancer Registries (NPCR) and/or the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program. Central registries may also be supported by state agencies, universities, and cancer centers. Participating central cancer registries include the following: Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Indiana, Iowa, Kentucky, Louisiana, Massachusetts, Maine, Maryland, Michigan, Mississippi, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Puerto Rico, Rhode Island, Seattle SEER Registry, South Carolina, Tennessee, Texas, Utah, Virginia, West Virginia, Wyoming. The authors assume full responsibility for analyses and interpretation of these data.

Kentucky: We would like to acknowledge the staff at the Kentucky Cancer Registry.

LCCS: We acknowledge the contributions of all who conducted this study which was originally reported as 10.1093/carcin/24.2.275.

NCCCS I & II: We would like to thank the study participants, and the NC Colorectal Cancer Study staff.

PLCO: The authors thank the PLCO Cancer Screening Trial screening center investigators and the staff from Information Management Services Inc and Westat Inc. Most importantly, we thank the study participants for their contributions that made this study possible. Cancer incidence data have been provided by the District of Columbia Cancer Registry, Georgia Cancer Registry, Hawaii Cancer Registry, Minnesota Cancer Surveillance System, Missouri Cancer Registry, Nevada Central Cancer Registry, Pennsylvania Cancer Registry, Texas Cancer Registry, Virginia Cancer Registry, and Wisconsin Cancer Reporting System. All are supported in part by funds from the Center for Disease Control and Prevention, National Program for Central Registries, local states or by the National Cancer Institute, Surveillance, Epidemiology, and End Results program. The results reported here and the conclusions derived are the sole responsibility of the authors.

SELECT: We thank the research and clinical staff at the sites that participated on SELECT study, without whom the trial would not have been successful. We are also grateful to the 35,533 dedicated men who participated in SELECT.

SLRCCS: We would like to thank Annika Lindblom and the SLRCCS Study staff and participants.

WHI: The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: <https://s3-us-west-2.amazonaws.com/www-whi-org/wp-content/uploads/WHI-Investigator-Long-List.pdf>

References

1. McAllister, K. *et al.* Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *Am J Epidemiol* **186**, 753-761 (2017).
2. Gauderman, W.J. *et al.* Update on the State of the Science for Analytical Methods for Gene-Environment Interactions. *Am J Epidemiol* **186**, 762-770 (2017).
3. Khera, A.V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224 (2018).
4. Zhang, X. *et al.* Circulating 25-hydroxyvitamin D and survival outcomes of colorectal cancer: evidence from population-based prospective cohorts and Mendelian randomisation. *Br J Cancer* **130**, 1585-1591 (2024).
5. Zhang, P. *et al.* Association of smoking and polygenic risk with the incidence of lung cancer: a prospective cohort study. *Br J Cancer* **126**, 1637-1646 (2022).
6. Tieu, S. *et al.* Genetic risk of type 2 diabetes modifies the association between lifestyle and glycemic health at 5 years postpartum among high-risk women. *BMJ Open Diabetes Res Care* **12**(2024).
7. Mooney, M.A. *et al.* Joint polygenic and environmental risks for childhood attention-deficit/hyperactivity disorder (ADHD) and ADHD symptom dimensions. *JCPP Adv* **3**, e12152 (2023).
8. Merino, J. *et al.* Interaction Between Type 2 Diabetes Prevention Strategies and Genetic Determinants of Coronary Artery Disease on Cardiometabolic Risk Factors. *Diabetes* **69**, 112-120 (2020).
9. Wang, Z., Shi, W., Carroll, R.J. & Chatterjee, N. Joint Modeling of Gene-Environment Correlations and Interactions Using Polygenic Risk Scores in Case-Control Studies. *Am J Epidemiol* (2024).
10. Kiyohara, C. & Yoshimasu, K. Genetic polymorphisms in the nucleotide excision repair pathway and lung cancer risk: a meta-analysis. *Int J Med Sci* **4**, 59-71 (2007).
11. Andersen, V. & Vogel, U. Interactions between meat intake and genetic variation in relation to colorectal cancer. *Genes Nutr* **10**, 448 (2015).
12. Darst, B.F. *et al.* Pathway-Specific Polygenic Risk Scores as Predictors of Amyloid-beta Deposition and Cognitive Function in a Sample at Increased Risk for Alzheimer's Disease. *J Alzheimers Dis* **55**, 473-484 (2017).
13. Goodman, M.O. *et al.* Pathway-Specific Polygenic Risk Scores Identify Obstructive Sleep Apnea-Related Pathways Differentially Moderating Genetic Susceptibility to Coronary Artery Disease. *Circ Genom Precis Med* **15**, e003535 (2022).
14. Choi, S.W. *et al.* PRSet: Pathway-based polygenic risk score analyses and software. *PLoS Genet* **19**, e1010624 (2023).
15. Pistis, G. *et al.* Gene set enrichment analysis of pathophysiological pathways highlights oxidative stress in psychosis. *Mol Psychiatry* **27**, 5135-5143 (2022).
16. Fernandez-Rozadilla, C. *et al.* Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nat Genet* **55**, 89-99 (2023).
17. Friis, S., Riis, A.H., Erichsen, R., Baron, J.A. & Sorensen, H.T. Low-Dose Aspirin or Nonsteroidal Anti-inflammatory Drug Use and Colorectal Cancer Risk: A Population-Based, Case-Control Study. *Ann Intern Med* **163**, 347-55 (2015).
18. Drew, D.A., Cao, Y. & Chan, A.T. Aspirin and colorectal cancer: the promise of precision chemoprevention. *Nat Rev Cancer* **16**, 173-86 (2016).
19. Drew, D.A. *et al.* Two genome-wide interaction loci modify the association of nonsteroidal anti-inflammatory drugs with colorectal cancer. *Sci Adv* **10**, eadk3121 (2024).

20. Liu, Z. *et al.* Annotation Query (AnnoQ): an integrated and interactive platform for large-scale genetic variant annotation. *Nucleic Acids Res* **50**, W57-W65 (2022).
21. Mi, H. & Thomas, P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol* **563**, 123-40 (2009).
22. Gauderman, W.J., Zhang, P., Morrison, J.L. & Lewinger, J.P. Finding novel genes by testing G × E interactions in a genome-wide association study. *Genet Epidemiol* **37**, 603-13 (2013).
23. Kawaguchi, E.S., Kim, A.E., Lewinger, J.P. & Gauderman, W.J. Improved two-step testing of genome-wide gene-environment interactions. *Genet Epidemiol* **47**, 152-166 (2023).
24. Kooperberg, C. & Leblanc, M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol* **32**, 255-63 (2008).
25. Zhang, P., Lewinger, J.P., Conti, D., Morrison, J.L. & Gauderman, W.J. Detecting Gene-Environment Interactions for a Quantitative Trait in a Genome-Wide Association Study. *Genet Epidemiol* **40**, 394-403 (2016).
26. Wang, Y. *et al.* TGF-beta1 mediates the effects of aspirin on colonic tumor cell proliferation and apoptosis. *Oncol Lett* **15**, 5903-5909 (2018).
27. Yan, M. *et al.* 15-Hydroxyprostaglandin dehydrogenase, a COX-2 oncogene antagonist, is a TGF-beta-induced suppressor of human gastrointestinal cancers. *Proc Natl Acad Sci U S A* **101**, 17468-73 (2004).
28. Burn, J. *et al.* Cancer prevention with aspirin in hereditary colorectal cancer (Lynch syndrome), 10-year follow-up and registry-based 20-year data in the CAPP2 study: a double-blind, randomised, placebo-controlled trial. *Lancet* **395**, 1855-1863 (2020).
29. Myung, S.J. *et al.* 15-Hydroxyprostaglandin dehydrogenase is an in vivo suppressor of colon tumorigenesis. *Proc Natl Acad Sci U S A* **103**, 12098-102 (2006).
30. Fink, S.P. *et al.* Aspirin and the risk of colorectal cancer in relation to the expression of 15-hydroxyprostaglandin dehydrogenase (HPGD). *Sci Transl Med* **6**, 233re2 (2014).
31. Kodach, L.L. *et al.* The effect of statins in colorectal cancer is mediated through the bone morphogenetic protein pathway. *Gastroenterology* **133**, 1272-81 (2007).
32. Fan, J. *et al.* Berberine and aspirin prevent traumatic heterotopic ossification by inhibition of BMP signalling pathway and osteogenic differentiation. *J Cell Mol Med* **27**, 3491-3502 (2023).
33. Fattahi, R., Mohebichamkhorami, F., Khani, M.M., Soleimani, M. & Hosseinzadeh, S. Aspirin effect on bone remodeling and skeletal regeneration: Review article. *Tissue Cell* **76**, 101753 (2022).
34. Dihlmann, S., Siermann, A. & von Knebel Doeberitz, M. The nonsteroidal anti-inflammatory drugs aspirin and indomethacin attenuate beta-catenin/TCF-4 signaling. *Oncogene* **20**, 645-53 (2001).
35. Szarynska, M., Olejniczak, A., Kobiela, J., Spychalski, P. & Kmiec, Z. Therapeutic strategies against cancer stem cells in human colorectal cancer. *Oncol Lett* **14**, 7653-7668 (2017).
36. Dihlmann, S., Klein, S. & Doeberitz Mv, M. Reduction of beta-catenin/T-cell transcription factor signaling by aspirin and indomethacin is caused by an increased stabilization of phosphorylated beta-catenin. *Mol Cancer Ther* **2**, 509-16 (2003).
37. Dunbar, K. *et al.* Aspirin Rescues Wnt-Driven Stem-like Phenotype in Human Intestinal Organoids and Increases the Wnt Antagonist Dickkopf-1. *Cell Mol Gastroenterol Hepatol* **11**, 465-489 (2021).
38. Chen, M., Wu, L., Zhan, H., Liu, T. & He, Y. Aspirin-induced long non-coding RNA suppresses colon cancer growth. *Transl Cancer Res* **10**, 2055-2069 (2021).
39. Lambert, S.A. *et al.* The Polygenic Score Catalog: new functionality and tools to enable FAIR research. *medRxiv* (2024).

40. Dai, J.Y., Kooperberg, C., Leblanc, M. & Prentice, R.L. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* **99**, 929-944 (2012).
41. Kawaguchi, E.S., Li, G., Lewinger, J.P. & Gauderman, W.J. Two-step hypothesis testing to detect gene-environment interactions in a genome-wide scan with a survival endpoint. *Stat Med* **41**, 1644-1657 (2022).
42. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774-781 (2023).
43. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res* **50**, D988-D995 (2022).
44. Li, W. *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* **49**, D1020-D1028 (2021).
45. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
46. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012).
47. McCarthy, D.J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Med* **6**, 26 (2014).
48. Mills, C. *et al.* PEREGRINE: A genome-wide prediction of enhancer to gene relationships supported by experimental evidence. *PLoS One* **15**, e0243791 (2020).
49. Thomas, P.D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci* **31**, 8-22 (2022).
50. Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**, 703-721 (2019).
51. Keller, M.C. Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol Psychiatry* **75**, 18-24 (2014).
52. Gauderman, W. & Morrison, J. Quanto 1.2.4: A computer program for power and sample size calculations for genetic-epidemiology studies, <https://keck.usc.edu/biostatistics/software/>. (2009).