Diet, inflammation, and genetic predictors of tissue-specific gene expression: A functionally-informed gene-environment interaction analysis for risk of colorectal cancer.

Paneen S. Petersen

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Ulrike Peters, Chair

Li Xu

Johanna Lampe

Polly Newcomb

Program Authorized to Offer Degree:

Epidemiology

University of Washington

Abstract

Diet, inflammation, and genetic predictors of tissue-specific gene expression: A functionally-informed gene-environment interaction analysis for risk of colorectal cancer.

Paneen S. Petersen

Chair of the Supervisory Committee:

Ulrike Peters

Department of Epidemiology

Colorectal cancer (CRC) is the second most commonly diagnosed cancer in both sexes combined worldwide, as well as the second leading cause of cancer deaths. Pathogenesis of CRC can be attributed to both genetic (G) and environmental (E) factors. Among environmental risk factors, dietary factors contribute significantly to the etiology of CRC. Extensive research has focused on relationships between dietary factors, food constituents, and dietary patterns to CRC risk. There is compelling evidence from observational studies which supports the carcinogenicity of increased consumption of processed meat, and red meat. Evidence for protective effects of fruit, vegetable, whole grains, or fiber intake is less certain. Dietary factors may also have a direct

effect on CRC risk by modifying the effect of genetic predisposition for CRC. Furthermore, evidence for a relationship between inflammation and cancer has also been established by previous epidemiological studies as well as through basic research, yet mechanisms that underlie the association between inflammation and CRC remain uncertain. C-reactive protein (CRP) is a positive acute-phase protein produced primarily in the liver as a response to infection, tissue injury, and systemic inflammation. Circulating levels of CRP are commonly used as a measure of chronic inflammation, and pre-diagnostic serum CRP has been associated with increased risk of CRC in previous studies. This dissertation research project utilized self-reported dietary histories and measures of circulating C-reactive protein to examine the interaction between these risk factors and genetic variation on the risk of colorectal cancer. Utilizing genotype array data from both population-based case-control studies and nested case-control studies in an international collaboration of well-characterized studies, the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), we employed a novel statistical approach to reduce multiple testing burden and increase power with *a priori* functional information from predicted tissue-specific gene expression. Our GxE analysis for dietary factors had a study sample comprised of 22 studies with 20,236 cases and 19,838 controls. In our dietary GxE analysis we detected a significant interaction ($P = 2.30 \times 10^{-7}$) between processed meat intake and variants predictive of RAC Family Small GTPase 1 (*RAC1*) gene expression in the colon and risk of CRC. The most significant interaction of a single variant within *RAC1* was rs2346263 ($\beta_{GxE} = -0.127$; $P = 1.32 \times 10^{-4}$). In our CRP GxE analysis of 5 studies comprising 5,423 individuals, we observed a 4 % increase of risk of CRC from a two-fold rise in CRP (mg/L) levels [OR:1.04 (95% CI:1.00, 1.08)]. There was significant heterogeneity in study specific effect sizes for the circulating CRP [$P = 7.6 \times 10^{-3}$]. At an FDR < 0.20 we detected no significant interactions

between circulating CRP and variants predictive of gene expression in the colon. Our results suggest that the *RAC1* gene may be involved in the ROS and inflammatory pathways that lead to CRC from intake of processed meat. Furthermore, our findings also suggest that CRP may not be associated with CRC and we did not find evidence that genetically determined gene expression influences the association between circulating CRP levels and CRC risk.

TABLE OF CONTENTS

LIST OF FIGURES ..... **ii**

LIST OF TABLES ..... **iii**

CHAPTER 1. INTRODUCTION **7**

CHAPTER 2. INTERACTIONS BETWEEN DIETARY INTAKE AND GENETIC PREDICTORS OF GENE EXPRESSION IN ASSOCIATION WITH RISK OF COLORECTAL CANCER **12**

2.1 ABSTRACT 12

2.2 INTRODUCTION 13

2.3 METHODS 14

2.4 RESULTS 18

2.5 DISCUSSION 20

CHAPTER 3. INTERACTIONS BETWEEN PREDICTED GENE EXPRESSION WITH CIRCULATING C-REACTIVE PROTEIN CONCENTRATION AND RISK OF COLORECTAL CANCER **48**

3.1 ABSTRACT 48

3.2 INTRODUCTION 50

3.3 METHODS 51

3.4 RESULTS 56

3.5 DISCUSSION 57

CHAPTER 4. CONCLUSION **74**

APPENDIX A: SUPPLEMENTAL TEXTS **75**

APPENDIX B: SUPPLEMENTAL FIGURES **103**

APPENDIX C: SUPPLEMENTAL TABLES **108**

LIST OF FIGURES

Figure 2.1 Association between red meat intake and risk of CRC. 28
Figure 2.2 Association between processed meat intake and risk of CRC. 29
Figure 2.3 Association between vegetable intake and risk of CRC. 30
Figure 2.4 Association between fruit intake and risk of CRC 31
Figure 2.5 Association between of dietary fiber intake and risk of CRC 32
Figure 2.6 Q-Q plots from gene interaction tests with sex- and -study specific quartiles of red meat intake for risk of CRC. 33
Figure 2.7 Q-Q plots from gene-set interaction tests with sex- and -study specific quartiles of processed meat intake for risk of CRC. 34
Figure 2.8 Q-Q plots from gene-set interaction tests with sex- and -study specific quartiles of vegetable intake for risk of CRC 35
Figure 2.9 Q-Q plots from gene-set interaction tests with sex- and -study specific quartiles of fruit intake for risk of CRC 36
Figure 2.10 Q-Q plots from gene-set interaction tests with sex- and -study specific quartiles of dietary fiber intake for risk of CRC 37
Figure 2.11 LD heat map for variants in *RAC1* gene-set 39
Figure 3.1 Association between circulating C-Reactive Protein (CRP) and risk of CRC 64
Figure 3.2 Q-Q plots from gene-set interaction tests with circulating CRP for risk of CRC 65
Figure 3.3 Q-Q plots from gene-set interaction tests with circulating CRP for risk of CRC excluding the Nurses' Health Study 66
Supplemental Figure 1. Flowchart of the data harmonization process (Chapter 2 & 3) 104
Supplemental Figure 2. Selection of CRP measurements from WHI sub-studies for analysis ... 106

LIST OF TABLES

Table 2.1 Characteristics of the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) 27
Table 2.2 MiSTi results for the Rac Family Small GTPase 1 (*RAC1)* gene-set and sex- and study- specific quartiles of processed meat intake. 34
Table 2.3 Generalized linear model regression results for *RAC1* interaction effects with processed meat intake. 38
Table 3.1 Characteristics of participants in GECCO with CRP measurements 62
Table 3.2 Assay characteristics of C-reactive Protein (CRP) data in GECCO 63
Supplemental Table 1. Genotyping platforms for participating studies (Chapter 2 & 3) 109
Supplemental Table 2. Missingness of dietary factor by outcome in GECCO 110
Supplemental Table 3. Characteristics of subjects in GECCO with CRP data by study 111
Supplemental Table 4. WHI studies with CRP measurements matched to subjects in GECCO……....113

ACKNOWLEDGEMENTS

I would like to thank my dissertation chair and academic advisor, Ulrike Peters, for her generous support and thoughtful mentorship. I am especially thankful for her critical input, encouragement, and advice throughout this dissertation project and my training. I am grateful for working with such an exceptional researcher and mentor.

My supervisory committee, including Li Hsu, Johanna Lampe, and Polly Newcomb, were generous with their time in providing their expertise and guidance for this dissertation project.

My parents deserve credit for wholeheartedly encouraging my scientific curiosity and nurturing a lifelong joy of discovery. My family also taught me about our responsibility to care for others in our community and it is for that purpose that I embarked on this program. Throughout, my family, friends and colleagues have been tireless cheerleaders and kept my spirit nourished with both kind words and niqipiaq.

I would also like to thank the Native organizations that have provided significant financial support of my graduate education because they support my vision of health research in the Alaska Native community led by Alaska Native researchers.

DEDICATION

For my Alaska Native community, because we have among the highest recorded incidence and mortality rates of colorectal cancer in the world. Piqpagigikpiñ.

CHAPTER 1. INTRODUCTION

Colorectal cancer (CRC) is the third most commonly diagnosed cancer in males and the second in females worldwide, with over 1.2 million new cancer cases and 600,000 deaths annually[1]. Many of the lifestyle and environmental risk factors associated with CRC risk are modifiable. Exposures associated with risk for CRC include obesity, tobacco use, consumption of red meat or processed meats, and excess alcohol consumption combined with a diet low in some micronutrients[2,3]. A decreased risk of CRC is associated with increased physical activity, dietary folate, dietary calcium, circulating levels of vitamin D, use of non-steroidal anti-inflammatory drugs (NSAIDs), and post-menopausal hormone therapy[2-5]. The role of the consumption of fruits and vegetables, and supplements such as vitamin D and vitamin B6 in risk of CRC is not as certain[6]. Although a relationship between inflammation and cancer has been established by previous epidemiological studies as well as through basic research, the mechanisms that underlie the relationship between inflammation and CRC remain unclear[7-9].

The advent of next-generation sequencing (NGS) offered the opportunity to use whole genome or whole exome sequencing with deep coverage (e.g. 30x) to uncover novel variation associated with CRC. New high penetrance mutations that are associated with CRC familial syndromes could be identified through linkage studies with whole genome (or whole exome) sequencing. Over fourteen genes associated with inherited syndromes associated with CRC have been discovered[10]. Inherited syndromes associated with CRC such as Lynch syndrome, familial adenomatous polyposis, MUTYH-associated polyposis, Peutz-Jeghers syndrome, juvenile polyposis and Cowden/PTEN hamartoma syndrome account for about 2% to 5% of all CRC cases[11]. Genome-wide association studies (GWAS) of sporadic CRC have identified over 100 susceptibility loci associated with CRC risk[10,12-15]. Association studies that have produced novel

loci were primarily performed within European populations, although some studies identified novel loci first in Asian study populations[16,17]. The majority of GWAS discovered loci over the last decade have been mapped to intronic or intergenic positions in non-coding regions, where the understanding of the putative consequences of this variation is still developing. Furthermore, the risk associated with each of these single nucleotide polymorphisms (SNPs) is very low and they explain only a fraction of genetic risk outside of familial syndromes[18]. Effect sizes of known CRC susceptibility loci are small, with ORs ranging from 1.04 to 1.57, and allele frequencies for the reported risk alleles range from 0.11% to 99.85%[10,12-15].

To fully understand the impact of risk factors on the etiology of CRC, it is important to examine whether interactions with genetic variants exist. Recently developed statistical methods offer increased power to examine gene-environment interactions (GxE) [22,25-28]. This has led to the discovery of novel loci with effects modified by environmental risk factors that would have been missed if only a GWAS screening approach for marginal effects had been used[19-21]. Previous GxE studies with dietary factors have identified significant interactions between loci and processed meat consumption[22], and alcohol[23]. Limited statistical power remains a primary concern in GxE analyses since the sample size required for detecting interactions is generally at least four times greater than that required for detecting a main effect of similar magnitude[24]. Using novel statistical methods such as set-based SNP testing with the incorporation of functional information can substantially reduce the multiple testing burden and increase statistical power to detect GxE interactions[25].

The focus of this research project was on interactions of dietary variables (intakes of red meat, processed meat, fruits, vegetables, and fiber) and C-reactive protein (CRP) with genetic variation and colorectal cancer risk. Our studies used genotype array data from both population-based

case-control studies and nested case-control from cohort studies in an international collaboration of well-characterized studies, the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO)[26]. This large consortium dataset allows for comprehensive "agnostic" genome-wide GxE interaction scans. Data for our exposures of interest were well-harmonized from self-reported dietary histories and pre-diagnostic CRP measurements collected within GECCO studies. The specific aims of this project were to test common SNPs for interactions with five dietary exposures (red meat, processed meat, vegetables, fruits and fiber) on the risk of CRC (Aim 1) and to test common SNPs for interactions with the pre-diagnostic level of circulating C-reactive protein on the risk of CRC (Aim 2). For both aims we implemented a novel mixed-effects set-based approach, the Mixed-Effects Score Test for Interactions (MiSTi)[25] for our GxE tests. Using MiSTi, we were able to incorporate *a priori* functional information from the genetically regulated gene-expression prediction tool, *PrediXcan*[27] based on transcriptome data from the Genotype-Tissue Expression (GTEx) Project[28].

REFERENCES

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin.* 2011;61(2):69-90.

2. Martinez ME. Primary prevention of colorectal cancer: lifestyle, nutrition, exercise. *Recent Results Cancer Res.* 2005;166:177-211.

3. Potter JD. Colorectal cancer: molecules and populations. *J Natl Cancer Inst.* 1999;91(11):916-932.

4. Limsui D, Vierkant RA, Tillmans LS, et al. Postmenopausal hormone therapy and colorectal cancer risk by molecularly defined subtypes among older women. *Gut.* 2012;61(9):1299-1305.

5. Chan AT, Ogino S, Giovannucci EL, Fuchs CS. Inflammatory markers are associated with risk of colorectal cancer and chemopreventive response to anti-inflammatory drugs. *Gastroenterology.* 2011;140(3):799-808, quiz e711.

6. Chan AT, Giovannucci EL. Primary prevention of colorectal cancer. *Gastroenterology.* 2010;138(6):2029-2043 e2010.

7. Terzic J, Grivennikov S, Karin E, Karin M. Inflammation and colon cancer. *Gastroenterology.* 2010;138(6):2101-2114 e2105.

8. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. *Cell.* 2010;140(6):883-899.

9. Coussens LM, Werb Z. Inflammation and cancer. *Nature.* 2002;420(6917):860-867.

10. Peters U, Bien S, Zubair N. Genetic architecture of colorectal cancer. *Gut.* 2015;64(10):1623-1636.

11. Ma H, Brosens LAA, Offerhaus GJA, Giardiello FM, de Leng WWJ, Montgomery EA. Pathology and genetics of hereditary colorectal cancer. *Pathology.* 2018;50(1):49-59.

12. Al-Tassan NA, Whiffin N, Hosking FJ, et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci Rep.* 2015;5:10442.

13. Schumacher FR, Schmit SL, Jiao S, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun.* 2015;6:7138.

14. Orlando G, Law PJ, Palin K, et al. Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Hum Mol Genet.* 2016;25(11):2349-2359.

15. Zeng C, Matsuda K, Jia WH, et al. Identification of Susceptibility Loci and Genes for

Colorectal Cancer Risk. *Gastroenterology.* 2016;150(7):1633-1645.

16. Jia WH, Zhang B, Matsuo K, et al. Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat Genet.* 2013;45(2):191-196.

17. Zhang B, Jia WH, Matsuda K, et al. Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet.* 2014;46(6):533-542.

18. Jiao S, Peters U, Berndt S, et al. Estimating the heritability of colorectal cancer. *Hum Mol Genet.* 2014;23(14):3898-3905.

19. Hancock DB, Soler Artigas M, Gharib SA, et al. Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLoS Genet.* 2012;8(12):e1003098.

20. Manning AK, Hivert MF, Scott RA, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet.* 2012;44(6):659-669.

21. Schoeps A, Rudolph A, Seibold P, et al. Identification of new genetic susceptibility loci for breast cancer through consideration of gene-environment interactions. *Genet Epidemiol.* 2014;38(1):84-93.

22. Figueiredo JC, Hsu L, Hutter CM, et al. Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet.* 2014;10(4):e1004228.

23. Gong J, Hutter CM, Newcomb PA, et al. Genome-Wide Interaction Analyses between Genetic Variants and Alcohol Consumption and Smoking for Risk of Colorectal Cancer. *PLoS Genet.* 2016;12(10):e1006296.

24. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol.* 1984;13(3):356-365.

25. Su YR, Di CZ, Hsu L, Genetics, Epidemiology of Colorectal Cancer C. A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics.* 2017;18(1):119-131.

26. Peters U, Jiao S, Schumacher FR, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology.* 2013;144(4):799-807 e724.

27. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet.* 2015;47(9):1091-1098.

28. Aguet F, Brown AA, Castel SE, et al. Genetic effects on gene expression across human tissues. *Nature.* 2017;550(7675):204-213.

CHAPTER 2. INTERACTIONS BETWEEN DIETARY INTAKE AND GENETIC PREDICTORS OF GENE EXPRESSION IN ASSOCIATION WITH RISK OF COLORECTAL CANCER

2.1 ABSTRACT

Background: Colorectal cancer (CRC) is a complex disease with many known genetic and environmental factors contributing to its development. Dietary factors play a significant role in CRC risk. We tested known dietary risk factors for interactions with genetic variation in a genome-wide gene-by-environment interaction (GxE) analysis. To improve power, we used a set-based approach with *a priori* functional information to select and aggregate variants for testing. *Methods:* Our analysis included 20,236 cases and 19,838 controls from 22 case-control and nested case-control studies. Dietary factors were modeled as study and sex-specific quartile increments of intake for processed meat, red meat, vegetables, fruits, and dietary fiber. We tested GxE interactions between dietary factors and 4,840 gene-sets of variants predictive of gene expression in the colon. We jointly tested fixed interaction effects of the weighted burden within each gene-set (burden) and residual GxE effects (variance) on risk of CRC with the Mixed Score Test for Interactions (MiSTi) approach using a Bonferroni corrected threshold of $P < 1 \times 10^{-5}$ for significance to adjust for multiple testing. *Results:* We observed a significant interaction between processed meat and the RAC Family Small GTPase 1 (*RAC1*) gene-set on the risk of CRC ($P = 2.30 \times 10^{-7}$). The most significant interaction of a single variant within *RAC1* was rs2346263 ($\beta_{GxE} = -0.127$; $P = 1.32 \times 10^{-4}$); however, the detected interaction between the *RAC1* gene-set and processed meat intake remained significant even after adjustment for rs2346263 ($P = 5.87 \times 10^{-6}$). *Conclusions:* Using functionally informed, set-based GxE testing we identified a novel interaction between processed meat intake and genetic predicted *RAC1* expression on CRC risk, which was not driven by a single variant. Processed meat intake may influence CRC

development through reactive oxygen species and inflammation pathways via genetically regulated *RAC1* expression in the colon. This finding supports that processed meat intake can influence CRC risk through interactions with common genetic variants.

2.2 Introduction

Colorectal cancer (CRC) is the second most commonly diagnosed cancer in both sexes combined worldwide, as well as the second leading cause of cancer deaths[1]. Pathogenesis of CRC can be attributed to both genetic (G) and environmental (E) factors[2]. Among environmental risk factors, dietary factors contribute significantly to the etiology of CRC[3]. Extensive research has focused on relationships between dietary factors, food constituents, and dietary patterns to CRC risk[4]. There is compelling evidence from observational studies which supports the carcinogenicity of increased consumption of processed meat, and red meat[5,6]. Evidence for protective effects of fruit, vegetable, whole grains, or fiber intake is less certain[6,7]. Dietary factors may also influence CRC risk by modifying the effect of genetic predisposition for CRC, which may explain partly the less consistent findings for some of the dietary factors. As diet is modifiable, the identification of gene-diet interactions has the potential for informing approaches to primary CRC prevention[8].

Genome-wide association studies (GWAS) of sporadic CRC have identified over 100 susceptibility loci that predict CRC risk[9-18]. Up to 35% of CRC is estimated to be attributable to genetic risk factors [19] yet the discovered loci only explain a fraction of heritable CRC risk[9-14,16,18,20]. Gene-environment interaction (GxE) studies of dietary factors could identify novel CRC susceptibility loci as well as identify biologic pathways and mechanisms of CRC etiology that are attributable to diet. Increasing our knowledge of the interplay between dietary and

genetic factors on CRC risk may also inform approaches to CRC prevention that address differential responses to dietary factors.

Previous genome-wide GxE studies of dietary factors and risk of CRC have identified a significant interaction with increased processed meat consumption and a single genetic variant [21]. Although the sample size required to detect a genome-wide GxE effect has been estimated at four times the sample size needed to detect a main genetic effect of similar magnitude[22], the integration of functional genomic information with novel statistical approaches can improve the power to detect interactions[23]. In this study, we have pooled a large sample of 43,951 participants from 22 studies in the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), the Colon Cancer Family Registry (CCFR), and the Colorectal Transdisciplinary (CORECT) Study. We tested whether genetic variation interacts with dietary intake of processed meat, red meat, fruit, vegetables, and fiber on the risk of CRC. To increase statistical power, we used a novel mixed effects set-based approach that incorporated a priori functional information to test for interactions between diet and variants predictive of gene expression in colon tissue based on data from the GTEx Project[24,25].

2.3 Methods

Study Participants

The study sample was comprised of 22 studies with 20,236 CRC or advanced adenoma cases (920; 4.5% of cases), and 19,838 controls of primarily European ancestry and Middle Eastern ancestry (Supplemental Text 1; and Table 2.1). All CRC cases were defined as colorectal adenocarcinoma and confirmed by medical records, pathologic reports, or death certificate. Advanced adenomas are clinically relevant precursors to CRC; cases were confirmed by medical

records, histopathology, or pathologic reports. Controls for advanced adenoma cases had a negative colonoscopy except for controls matched to cases with distal adenoma, which either had a negative sigmoidoscopy or colonoscopy exam. All studies received ethical approval by their respective Institutional Review Boards.

Genotype Data

Genotype data was generated from germline DNA on several platforms and pooled into two data sets, genotyping platform details are in Supplementary Table 1. Sample collection, genotyping, and quality control (QC) including average sample and variant call rates, and concordance rates for blinded duplicates have been previously described in detail for all GWAS data[2,12,14,18]. To summarize, we excluded duplicates, close relatives defined as individuals that are second degree or more closely related, and samples with discrepancies between reported and genotypic sex, and samples with call rate <98%. Directly genotyped variants were excluded for a call rate <98%, Hardy-Weinberg Equilibrium (HWE) with $P < 1 \times 10^{-4}$ in controls, and an effect allele count less than 3. After QC, genotype data was imputed to the Haplotype Reference Consortium (HRC version r1.0) using the Michigan Imputation Server with phasing option set to ShapeIT v2.r790[26-28]. Variants were restricted to imputation accuracy $R2 > 0.3$. We performed principal components analysis (PCA) using PLINK 2.0[29] and analyses were restricted to samples that clustered with European descent in PCA. Variants were annotated using the human assembly GRCh37 (h19).

Functional Information

Variant weights from the *PrediXcan*[30] approach for estimating genetically determined gene expression were used as a priori functional information for set-based GxE testing. We downloaded the tissue-specific gene expression models of the *PrediXcan* approach from the

publicly accessible PredictDB Data Repository (http://hakyimlab.org/predictdb/). Gene expression prediction models for colon sigmoid and colon transverse tissues were built using data from the GTEx Project V6p[25] of predominantly European ancestry samples (n=169). The *PrediXcan* reference cohort, datasets, and model building have previously been described in detail[30]. Briefly, Gamazon et al. developed additive models of gene expression levels in *PrediXcan* using jointly measured genome-wide genotype data and RNA-seq data. We excluded the colon sigmoid *PrediXcan* models from our analyses because sample attributes available from the GTEx Portal indicated that the sampling target for the sigmoid colon was muscularis and not mucosa, which is the tissue of interest for CRC (GTEx_Data_V6_Annotations_SampleAttributesDS.txt). Samples for the GTEx transverse colon reference data included the entire colonic wall, including mucosa. We selected the transverse colon (hereafter referred to as "colon" in this paper) prediction models trained with genotype data imputed to the 1000 Genome Project v3[31]. The models were comprised of 4,878 gene sets with 145,258 unique variants. Of those, 99.2% of genes and 86.9% of variants were covered in our genotype data. Variants in the *PrediXcan* models that did not intersect with our genotype data were excluded from analyses.

Epidemiological Data

Basic demographics, lifestyle and dietary risk factors were collected through self-report using in-person interviews and/or structured questionnaires. Harmonization of individual level data was performed at the GECCO data coordinating center. Data harmonization comprised a multi-step process for data reconciliation of each study's unique protocols and data-collection instruments (Supplemental Figure 1). All variables for analysis were collected at study referent time, which was defined as study enrollment or blood collection for cohort studies and one to two years

before sample ascertainment for case-control studies to ensure exposures were assessed before cancer diagnoses. Age at referent time was defined in years and modeled continuously.

Dietary factors were assessed using food frequency questionnaires (FFQs) or diet history[21,32-36]. The dietary variables for fruits, vegetables, and red or processed meats were measured in servings/day and directly calculated from FFQ or diet history. Because fiber intake is a component of food items recorded, participating studies derived total fiber intake (g/day) using reference databases that define the fiber components of food items. Dietary variables were coded as sex- and study-specific quartiles with quartile cutoffs determined within the controls of each study and sex. Some studies had less variation in red and processed meat intakes, largely driven by fewer questions in those categories. In those instances, we assigned intake in those studies to only 2nd and 3rd quartiles of the variables. Total energy intake was calculated in kcal/day and modeled as a scaled continuous variable. Five studies did not have total energy intake available as part of their study design (Supplemental Table 2), so we used an indicator (1/0) variable to account for whether the value for total energy was missing.

Statistical Analyses

Statistical analyses of all data were conducted centrally at the GECCO coordinating center to ensure a consistent analytical approach. All analyses were conducted using R[37]. Multivariable logistic regression was used to estimate study-specific odds ratios (ORs) and 95% confidence intervals (CIs) for the association between each dietary factor and CRC risk; study-specific estimates were then combined using fixed-effects meta-analysis. We used the set-based approach Mixed Effects Score Test for Interactions (*MiSTi*) to test for gene-diet interactions[23]. *MiSTi* provides a unified hierarchical regression framework for modeling GxE effects that has two components: interaction of E with genetically predicted gene expression as fixed effects (burden

component), and residual heterogeneous GxE effects as random effects (variance component). We used the colon *PrediXcan* variant weights to calculate the genetically predicted gene expression. We minimally adjusted the *MiSTi* model for age at the referent time, sex, smoking, study, genotyping platform, genotyping phase, total energy intake and principal components (PC's) to account for potential population substructure. We used Fisher's combination method[38] to calculate an overall p-value combining the p-values from both the fixed and random effects. We assessed the potential inflation due to unaccounted for confounders by using quantile-quantile (Q-Q) plots of the p-values for the fixed effects and random effects. We also calculated genomic control (λ). Of 4878 gene-sets, 38 had a small number of variants (1 to 4) which created too sparse of data cells for testing with *MiSTi*, and thus excluded. We used Bonferroni correction to account for multiple comparisons, dividing 0.05 by the number of GxE models in MiSTi ($0.05/4840 = 1.03 \times 10^{-5}$) to determine a p-value threshold for significant gene-diet interactions. For any genes with significant interactions fit a GLM GxE regression for each variant, adjusting for age at the referent time, sex, smoking, study, genotyping platform, genotyping phase, total energy intake and principal components (PC's). To assess whether a significant gene-set interaction was driven by a single variant, we repeated the MiSTi for the gene-set with adjustment for the most significant variant ($P< 0.05/n$ variants) and its GxE. .

2.4 RESULTS

Characteristics of the 20,236 cases and 19,838 controls from 22 studies included in our gene-diet interaction analyses are summarized in Table 2.2. There were a similar number of men and women. Study-specific mean age ranged from 55.4 to 69.7 years. Per quartile increase in red meat intake was associated with 12% greater risk of CRC [OR:1.12 (95% CI:1.10, 1.14)] (Figure 2.1). Processed meat intake had a similar effect with 10% increased risk of CRC per quartile

increase of processed meat intake [OR:1.10 (95% CI:1.07, 1.13)] (Figure 2.2). A decreased risk for CRC was associated with per quartile increase of vegetable intake [OR: 0.90 (95% CI: 0.89, 0.92)] (Figure 2.3), fruit intake [OR: 0.90 (95% CI:0.88, 0.92)] (Figure 2.4), and dietary fiber intake [OR: 0.85 (95% CI:0.82, 0.87)] (Figure 2.5). Tests for heterogeneity in the fixed effect meta-analyses for the association between each dietary factor and CRC risk indicated statistically significant variation in effect sizes between studies for all evaluated dietary factors with I^2 values ranged from 59.3% to 73.2%.

In total, 4,840 gene-based variant sets predictive of gene expression in the colon were tested using the *MiSTi.* All gene sets had R^2 values for predicted gene expression greater than 0.01 (median = 0.06; range from 0.02 to 0.70). The number of variants in each gene set ranged from 1 to 227, with a median of 25 variants. Q-Q plots of p-values for the interaction tests from *MiSTi* (Burden, Variance, and Fisher's Combination) did not indicate inflation from the null distribution overall for any of the dietary factors (Figure 2.6 to Figure 2.10), consistent with the genomic control values (λ range: 1.02 to 1.13).

There was a statistically significant interaction detected between the gene RAC Family Small GTPase 1 (*RAC1*) and processed meat from the *MiSTi* Fisher combination test (Fisher's $_{\text{Combined}}$, $P = 2.30 \times 10^{-7}$, Figure 2.7). The signal for the interaction between processed meat and genetically predicted gene expression for the *RAC1* gene set ($P = 7.37 \times 10^{-5}$) was slightly stronger than the signal from the interaction between processed meat and residual effects for the *RAC1* gene set ($P = 2.01 \times 10^{-4}$), but clearly both components contributed to the interaction. No additional statistically significant interactions were detected between the analyzed genes and any of the other dietary factors after Bonferroni correction for multiple testing.

There were 24 variants in the *RAC1* gene set, with an R^2 of 0.13 for variation of *RAC1* expression in the colon. Characteristics of the variants in the *RAC1* gene set are described in Table 2.2. The effect allele frequency of variants in the *RAC1* set ranged from 0.04 to 0.73. A heatmap plot of R^2 values for linkage disequilibrium (LD) among all variants in the *RAC1* gene set (Figure 2.11) showed many of the variants were highly correlated. In the GLM GxE regression variant rs2346263 showed and the most significant interaction (β^{GxE} = - 0.127; P =1.32 x 10^{-4}) on CRC risk (Table 2.4). The rs2346263 effect allele is common (allele frequency = 0.29) and very weakly correlated ($R^2 \leq 0.02$) with 23 other variants in the gene set (Figure 2.11). A test of the *RAC1* x processed meat interaction with the *MiSTi* using a model that included terms for both rs2346263 genotype and the interaction between rs2346263 genotype and processed meat intake did not substantially alter the magnitude of signal for the interaction between processed meat and genetically predicted gene expression for the *RAC1* gene set (Burden, P =6.71 x 10^{-5}). However, the adjustment for rs2346263 in *MiSTi* weakens the interaction between processed meat and residual effects for the *RAC1* gene set (P = 1.63 x 10^{-2}).

2.5 Discussion

We observed strong associations between red meat and processed meat intakes with CRC risk, and a protective effect of vegetable, fruit, and fiber intakes with CRC risk. These results are consistent with previous evidence from observational studies although the protective effect of vegetables, fruits, and fiber has been less supported in the literature[39]. Using our powerful statistical approach *MiSTi* we observed a statistically significant interaction between RAC Family Small GTPase 1 (*RAC1*) and processed meat intake on risk of CRC. We did not observe statistically significant interaction effects between predicted gene expression and intake of any red meat, fruit, vegetable, or fiber on risk of CRC.

The *MiSTi* approach allowed us to examine signals for both interaction effects from predicted gene expression and residual GxE effects and we found that both contributed to the observed interaction between processed meat and *RAC1* on CRC risk. In exploratory analysis of individual variant interaction effects between the *RAC1* gene-set and processed meat, no single variant was a primary driver of the observed interaction, although the signal from residual GxE effects between variants in the *RAC1* gene set and processed meat may have been partially driven by rs2346263. Rs2346263 is an intronic variant located in the Zinc Finger Protein 853 (ZNF853) gene at 7p22.1, and this locus has not previously been associated with CRC risk. However, while this variant is located in *ZNF853*, in the *PrediXcan RAC1* expression model for the transverse colon, the presence of at least one copy of the effect allele rs2346263 is expected to contribute to increased *RAC1* expression. Our interpretation of these results is that the primary driver of the *RAC1* gene-set interaction with processed meat is the genetically predicted *RAC1* expression in the colon.

Increasing evidence shows a carcinogenic effect of processed meat on CRC. Processed meats may be cured, smoked, or cooked and include ready-to-cook and ready-to-eat products mostly made from red meat (mammalian muscle meat); these may also contain other meat products such as poultry, offal or meat by-products. Carcinogenic substances hypothesized to contribute to the increased risk of CRC from red and processed meat intakes include heterocyclic amines (HCAs)[40], polycyclic aromatic hydrocarbons (PAHs)[41], N-Nitroso-Compounds (NOC)[42]; and heme iron[43]. Of the potential mechanisms, it is perhaps the exogenous and endogenous formation of NOC that is uniquely enriched in processed meats[44]. The primary source of endogenous NOC exposure is from the metabolism of ingested nitrate or nitrite by bacteria and activated macrophages in the gut. While a significant dietary source of nitrate is vegetables, the

endogenous formation of NOCs from vegetable intake is inhibited by substances also abundant in vegetables: vitamins C and E, and polyphenols[45]. In contrast, the endogenous NOC formation of nitrate and nitrite from processed meat intake is catalyzed by amines, amides, and heme present in the meats[46]. While not specific to only processed meat, heme iron promotes the formation of NOCs, particularly in the colon where heme iron can accumulate because of poor absorption in the small intestine[47]. Nitrosyl heme found in processed meat is thought to be more cytotoxic than heme in unprocessed red meat[43]. Exogenously formed NOCs from nitrite or nitrate additives used in the curing process that can be enhanced by high-temperature cooking are also elevated in processed meats. Mixing with additional subcutaneous fat is common in the processing of meat products, thus it typically contains more fat than red meat[48] which likely increases production of lipid oxidation products (LOP). Dietary lipids have pro-inflammatory and pro-fibrogenic effects within the colon, with considerable amounts of exogenous and endogenous LOP that become available at the level of the colon mucosa. As heme iron accumulates in the colon both heme iron and NOC formation stimulate production of reactive oxygen species (ROS) which in turn also increases lipid peroxidation[43,49]. As a result, processed meat intake leads to higher exposure levels of NOC and LOP than unprocessed meat. The hyperproliferation of epithelial cells due to cytotoxic effects from LOP and heme can lead to accumulation of mutations. NOCs provoke several DNA mutations associated with CRC, and red meat is directly associated in a dose dependent fashion with the colonic formation of O6-carboxymethyl guanine, a NOC-specific DNA adduct associated with CRC[50]. Heme iron intake has also been associated with G > A transition in the *APC* gene and thought to have a key role in enhancing the genotoxic effects of NOC-stimulated DNA alkylation[51]. Carcinogenic effects in the colon from lipid peroxidation leads to oxidative stress and the creation of reactive aldehydes

and formation of DNA adducts associated with the development of cancer[52,53]. LOP induced specifically by heme iron are associated with increased markers of colonic inflammation, genotoxicity, and permeability *in vivo*[54]. In summary, processed meat my impact CRC risk through the accumulation of hem and formation of NOCs which lead to increase in ROS and LOPs production all of which can increase DNA mutation and adducts.

RAC1 is a member of the family of Rho GTPases genes with a pro-inflammatory effect and it is connected to a network of proteins integral to the regulation of immune response, and consequently inflammation[55]. *RAC1* regulates neutrophil functions and stimulates NADPH oxidase activity in macrophages, and *RAC1* activity is associated with inflammatory response in IBD patients[56]. Functions of the *RAC1* GTPase are regulated by cycling between an inactive GDP-bound form and an active GTP-bound form. *RAC1* regulates many intracellular signaling pathways involved in tumorigenesis, invasion, and metastasis including mTOR, NF-κB, JNK, and ROS[57]. There is also an alternatively spliced variant of *RAC1*, *RAC1b*, that is a self-activating GTPase found predominantly in the GTP-bound form[58]. *RAC1b* has more enhanced binding to proteins involved in transcriptional regulation, cell-cell adhesion, and motility than *RAC1*[59]. *RAC1b* selectively promotes NF-κB activation and signaling and has been shown to induce cell cycle progression and cancer cell survival in colorectal tumors[60].

RAC1 expression in normal tissues is ubiquitous, however dysregulated *RAC1* expression has been described in studies comparing cancerous tissue to normal tissue for melanoma, lung, breast, brain and colon cancers[57,61] and *RAC1* inhibitors have been evaluated as therapeutic targets for cancer[57,62]. *RAC1* -driven ROS production and NF-κB activation are thought to be mediators in the process of CRC initiation[63]. The role of *RAC1* in colorectal tumor progression has also been investigated in vivo by genetic modification of the human colorectal

adenocarcinoma cell line SW620 showing that overexpression of *RAC1* resulted in accelerated tumorigenic process while inhibition of *RAC1* completely suppressed tumor formation[64]. *RAC1* has also been identified as an important regulator of tumor-initiating intestinal stem cells through ROS production and NF-κB activation in *WNT*-driven tumorigenesis after loss of *APC*[63]. In summary, there is strong biological support for a role of *RAC1* in colorectal carcinogenesis and evidence of *RAC1* involvement with mechanisms speculated to underly the increased risk of CRC from processed meat, namely through impact on ROS on CRC. DNA mutations triggered by the elevated dietary NOCs and heme iron could initiate CRC tumors, followed by a tumor progression enhanced by the inflammation process associated with oxidative stress and lipid peroxidation products. The plausible biological explanation for our observed interaction between *RAC1* gene expression and increased processed meat intake would then be that detrimental effects from increased consumption of processed meat are mediated by *RAC1* involvement in ROS and inflammation pathways.

Our study had several strengths. This the largest study to date of gene-diet interactions on CRC risk and the first to incorporate tissue-specific functional information that prioritized variants predictive of gene expression in colon tissue. Our pooled sample of 20,236 cases and 19,838 controls comprised of well harmonized dietary risk factors as well as genome-wide genotype data enhanced by imputation. The harmonization of data at the individual level allowed us to reduce the impact of inter-study heterogeneity and sample outliers. We also employed a novel set-based testing approach with increased power to detect interaction effects on risk of CRC through the aggregation of variants into biologically relevant sets. This approach also allowed us to jointly test dietary factors for interactions between predicted gene expression in the colon and random heterogenous interaction effects.

There are also some limitations to our analysis. The functional information used in our GxE tests was suboptimal because sampling for GTEx tissue specimens in the sigmoid and transverse colon tissues was not specific to the most relevant tissues to CRC tumorigenesis: colorectal mucosa, or more specific intestinal stem cells[65]. Targets for GTEx colon tissue sampling were muscularis in the sigmoid and the entire colonic wall in the transverse colon. Therefore, we did not use the prediction models for gene expression in the sigmoid colon in our analyses. An ideal transcriptome reference set would have been derived from samples across the colon and rectum that measured expression in epithelial tissue of the mucosa, or more precisely from stem cells in colonic crypts where tumor development originates. We did not exclude advanced adenoma cases from studies in our analysis, as they are important preclinical precursors of CRC with overlapping risk profiles for both genetic and environmental factors[66]. Heterogeneity in the estimated associations between dietary factors and CRC within GECCO may have been introduced by misclassification from the harmonization of dietary factors of varied composition into categorical servings across GECCO. We did not divide our large sample into separate discovery and replication sets, as a pooled analysis is the most powerful approach and the sample size required for detecting interactions is much greater than that required for detecting a main effect of similar magnitude. We believe the choice for a combined analysis across all studies was the most powerful analytical approach. Lastly, we were not able to explore if our finding generalizes to other ancestral groups, however, our findings from this study will be informative to colleagues with non-European study populations.

Future analysis using this GxE approach with predicted gene expression would be improved by models trained on a larger reference set of transcriptome data specific to colon mucosa and the stem cell niche. Additionally, alternative splicing isoforms are integral to proteome complexity

and diversity, and aberrant splicing has been associated with the initiation and survival of tumors[67]. Distinguishing between isoforms in predicted gene expression in *a priori* information could further clarify mechanisms by which *RAC1* modifies the association between intake of processed meat and CRC risk. Future studies should also attempt replication of this finding in independent populations and functional analysis to identify specific causative genetic variants.

In conclusion, our study was the first to analyze gene-diet interactions using a set-based approach with variants predictive of tissue-specific gene expression in the colon. With a large GWAS consortium built on well-characterized studies, we were well-positioned to identify potential interactions between genes and dietary risk factors with respect to CRC risk. We identified a novel interaction between the *RAC1* gene and processed meat intake that supports processed meat consumption can influence CRC development through ROS and inflammatory pathways through common genetic variants.

Table 2.1 Characteristics of the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO)

Study Name	Study Design	Cases[a]	Controls	Female %	Mean Age (SD)
Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC)[68,69]	clinical trial	137	28	0	57.2 (4.7)
Cancer Prevention Study II (CPS II)[70]	cohort	490	498	50.2	68.8 (5.4)
Colon Cancer Family Registry (CCFR)[71,72]	case-control	3,238	2,530	51.3	55.4 (11.8)
Colorectal Cancer Genetics & Genomics (CRCGEN)	case-control	705	827	41.6	66.1 (11.3)
Darmkrebs: Chancen der Verhütung durch Screening (DACHS)[52,73]	case-control	2,371	2,200	39.9	68.7 (10.4)
Diet, Activity and Lifestyle Study (DALS)[74]	case-control	1,104	1,170	44.9	63.9 (9.9)
French Association Study Evaluating RISK for sporadic colorectal cancer (ASTERISK)[75]	case-control	877	943	41.6	65.2 (10.6)
Hawaiian Population-Based Case-Control Study (COLO2&3)[76]	case-control	87	121	44.7	65.4 (11.3)
Health Professionals Follow-up Study (HPFS)[77]	cohort	485	571	0	62.7 (8.9)
Kentucky Case-Control Study (Kentucky)[78,79]	case-control	929	1,026	51.8	62.8 (9.5)
Melbourne Collaborative Cohort Study (MCCS)[80,81]	cohort	752	671	48.1	59.6 (7.6)
Molecular Epidemiology of Colorectal Cancer (MECC)[82]	case-control	4,087	3,159	48.2	69.7 (12)
Multiethnic Cohort (MEC)[83]	cohort	312	334	46.9	63 (8)
Newfoundland Case-Control Study (NFCCR)[84]	case-control	184	457	41.2	60 (9)
Nurses' Health Study (NHS)[85]	cohort	811	1,351	100	58.4 (6.8)
Ontario Familial Colorectal Cancer Registry (OFCCR)[86]	case-control	498	518	53.6	62.3 (7.7)
Physicians' Health Study (PHS)[87,88]	cohort	373	389	0	58.9 (9)
Postmenopausal Hormone Study - Colon Cancer Family Registry (PMH-CCFR)[89]	case-control	276	122	100	62.7 (7.1)
Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)[90-92]	cohort	421	640	38.3	64.5 (5.1)
Swedish Mammography Cohort and Cohort of Swedish Men (SMC_COSM)[93]	cohort	419	495	60.1	63.7 (8.5)
VITamins And Lifestyle cohort (VITAL)[94]	cohort	249	262	45.8	66.2 (6.2)
Women's Health Initiative (WHI)[95-97]	cohort	1,431	1,526	100	66.4 (6.6)
Overall		20,236	19,838	51.9	64.0 (11.0)

a. Health Professionals Follow-up Study (HPFS) and Nurses' Health Study (NHS) cases include 312 and 513 advanced adenoma cases, respectively.

Figure 2.1 Association between red meat intake and risk of CRC. Odds Ratio estimates correspond to each quartile increase in servings/day of red meat, adjusted for age at the referent time, sex, smoking (PHS only), and total energy intake.

Sex- and study- specific quartiles of red meat (servings/day) and risk of CRC

Study	Design	Country	Cases	Controls	OR (95% CI)
ASTERISK	Case-control	France	860	938	1.32 (1.01,1.74)
ATBC	Clinical Trial	Finland	137	28	1.32 (0.90,1.94)
CCFR	Case-control	USA, Australia	2,844	2,337	1.09 (1.04,1.14)
Colo23	Case-control	USA	87	121	0.98 (0.43,2.21)
CPSII	Cohort	USA	490	498	1.11 (0.99,1.25)
CRCGEN	Case-control	Spain	705	827	1.03 (0.94,1.14)
DACHS	Case-control	Germany	2,370	2,192	1.26 (1.18,1.34)
DALS	Case-control	USA	1,104	1,170	1.13 (1.03,1.23)
HPFS	Cohort	USA	485	571	1.22 (1.07,1.38)
Kentucky	Case-control	USA	918	1,009	1.26 (1.15,1.38)
MCCS	Cohort	Australia	752	671	1.03 (0.93,1.14)
MEC	Cohort	USA	312	334	1.06 (0.90,1.25)
MECC	Case-control	Israel	4,070	3,158	1.13 (1.08,1.18)
NFCCR	Case-control	Canada	180	452	0.97 (0.84,1.13)
NHS	Cohort	USA	811	1,351	1.09 (0.99,1.19)
OFCCR	Case-control	Canada	476	497	1.30 (1.16,1.45)
PHS	Cohort	USA	367	380	0.98 (0.85,1.12)
PLCO	Cohort	USA	421	640	1.12 (0.98,1.28)
PMH-CCFR	Case-control	USA	265	113	1.10 (0.87,1.38)
SMC_COSM	Cohort	Sweden	415	492	1.04 (0.91,1.18)
VITAL	Cohort	USA	249	262	1.08 (0.91,1.29)
WHI	Cohort	USA	1,429	1,523	1.01 (0.94,1.09)
Meta-analysis			**19,747**	**19,564**	**1.12 (1.10,1.14)**

Heterogeneity: $I^2 = 58.9\%$ P = 1.13e-04

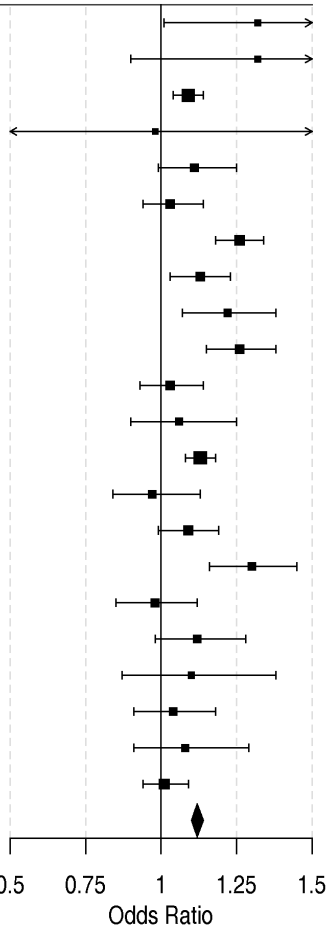


Figure 2.2 Association between processed meat intake and risk of CRC. Odds Ratio(OR) estimates correspond to each quartile increase in servings/day of processed meat, adjusted for age at the referent time, sex, smoking (PHS only), and total energy intake.

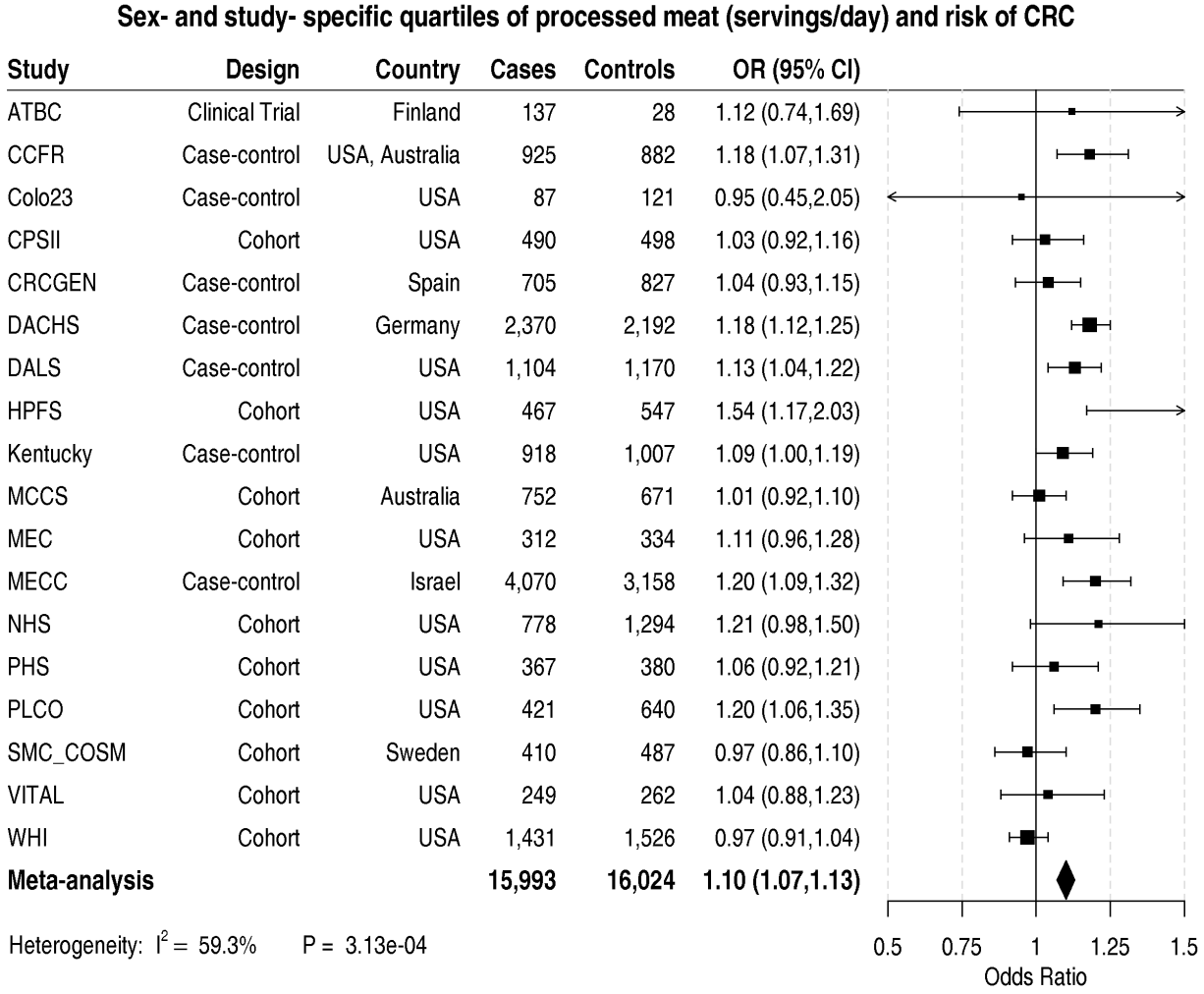


Figure 2.3 Association between vegetable intake and risk of CRC. Odds Ratio(OR) estimates correspond to each quartile increase in servings/day of vegetables, adjusted for age at the referent time, sex, smoking (PHS only), and total energy intake.

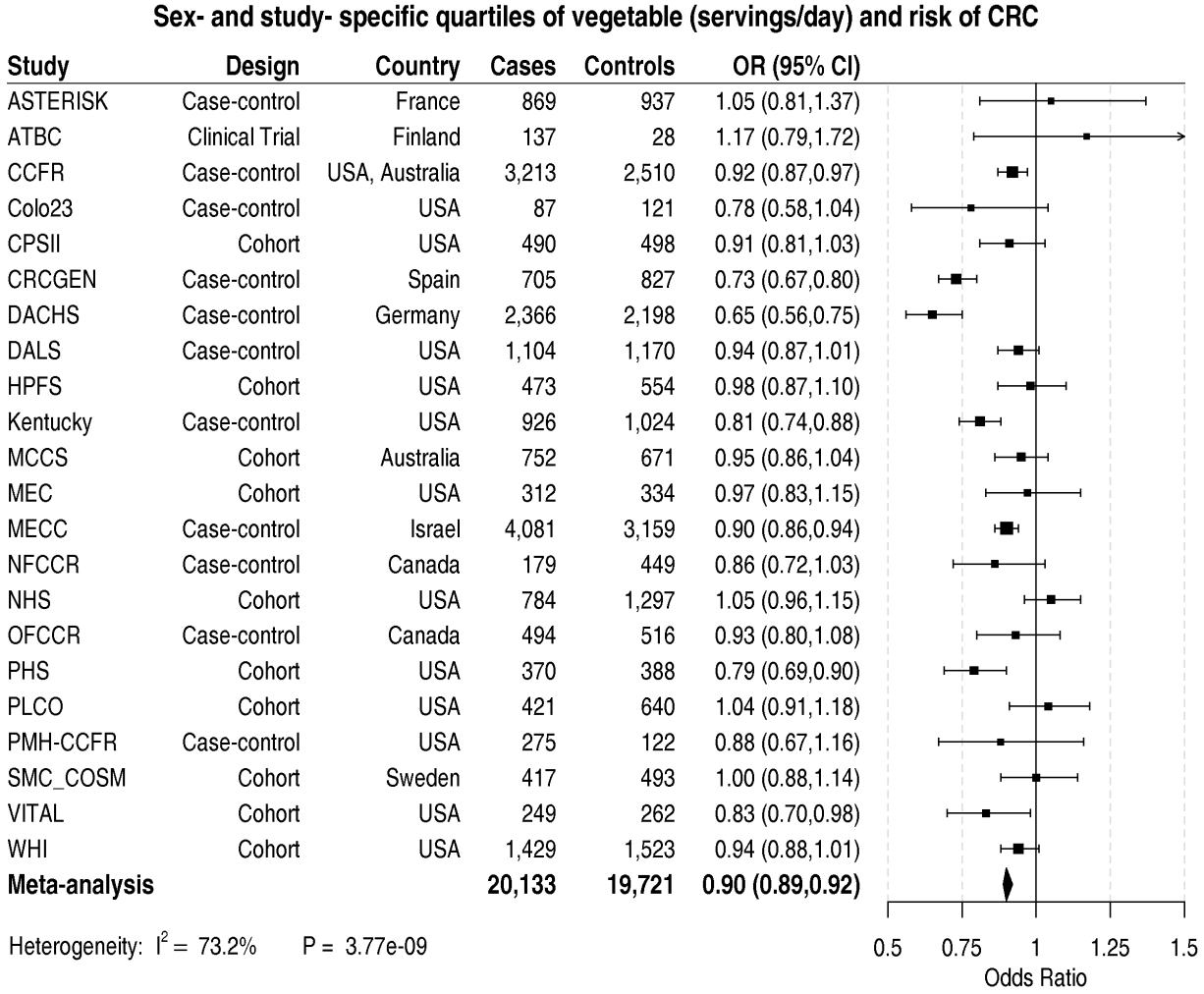


Figure 2.4 Association between fruit intake and risk of CRC. Odds Ratio(OR) estimates correspond to each quartile increase in servings/day of fruit, adjusted for age at the referent time, sex, smoking (PHS only), and total energy intake.

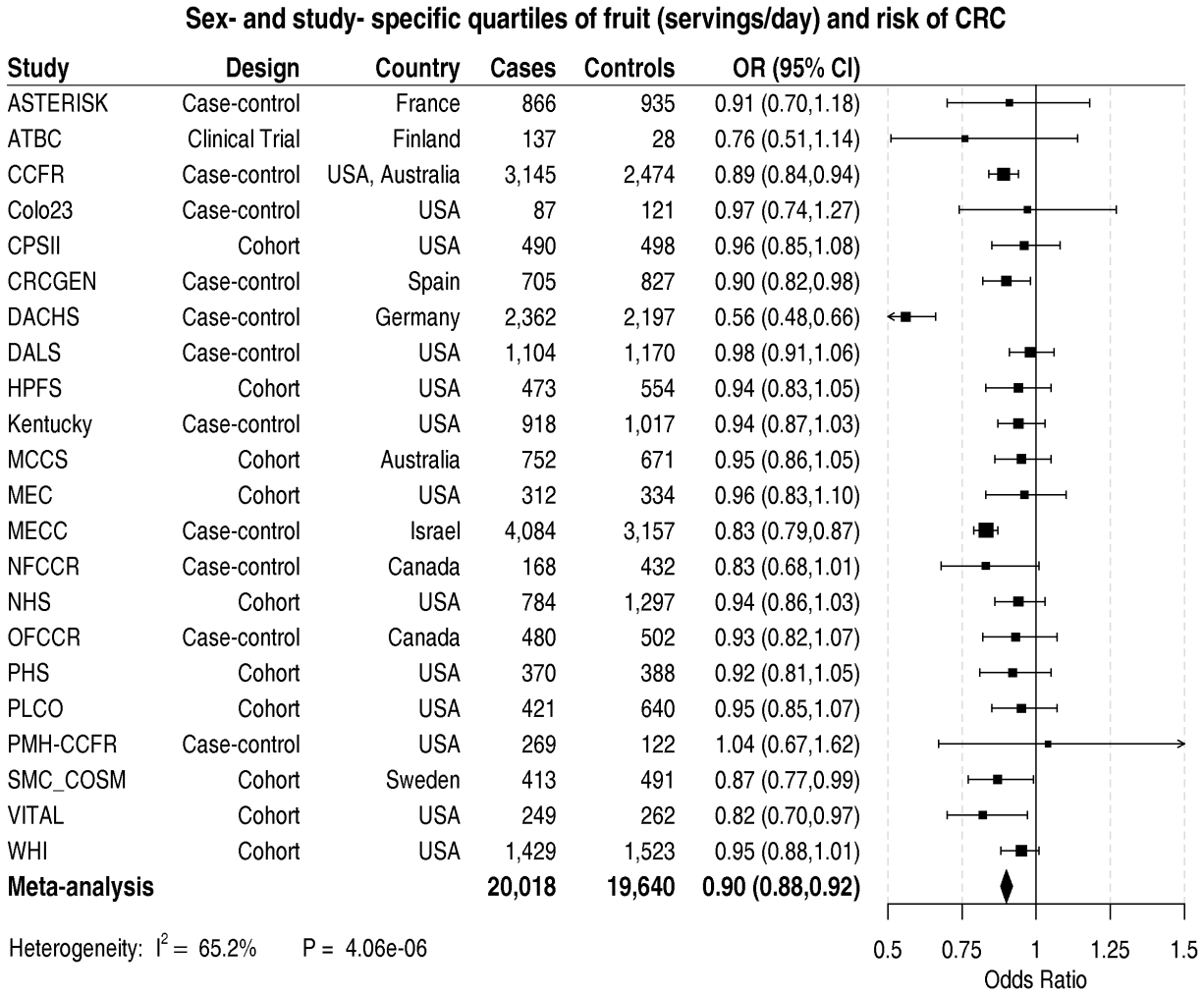


Figure 2.5 Association between of dietary fiber intake and risk of CRC. Odds Ratio(OR) estimates correspond to each quartile increase in g/day of fiber, adjusted for age at the referent time, sex, smoking (PHS only), and total energy intake.

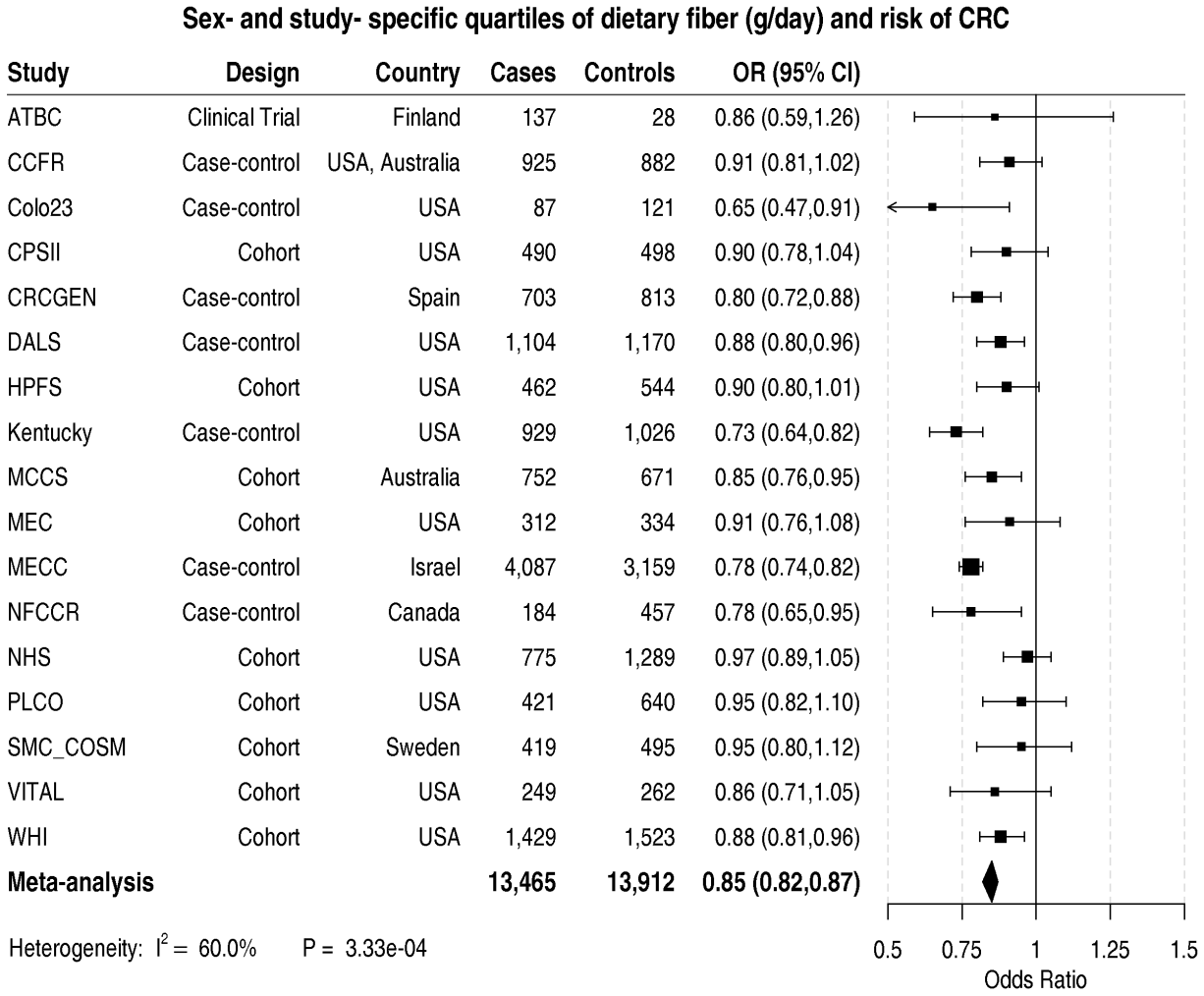


Figure 2.6 Q-Q plots from gene interaction tests (n=4840) with sex- and -study specific quartiles of red meat intake for risk of CRC. Results are from the MiSTi burden component, MiSTi variance component, and combined MiSTi burden and variance components using Fisher's Combination method. MiSTi models were adjusted for age at the reference time, sex, study, genotyping phase, total energy consumption, and the principal components to account for potential population substructure. The Bonferroni corrected threshold of $P = 1.03 \times 10^{-5}$ is labeled in red.

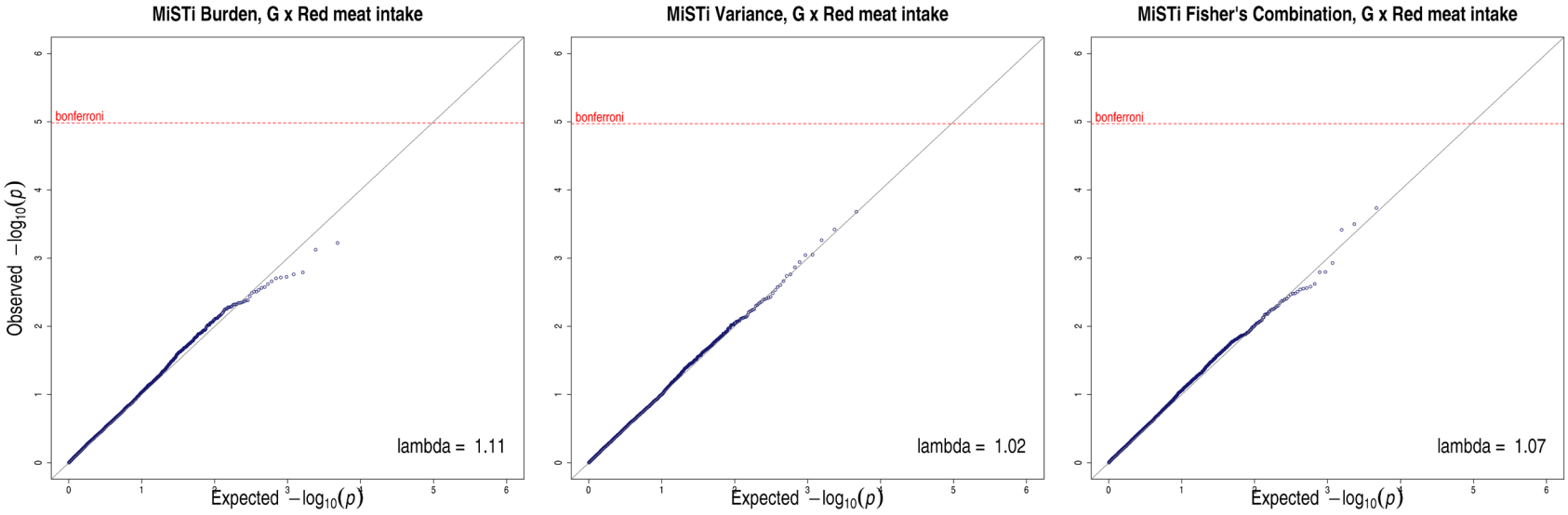


Figure 2.7 Q-Q plots from gene-set interaction tests (n=4840) with sex- and -study specific quartiles of processed meat intake for risk of CRC. Results are from the MiSTi burden component, MiSTi variance component, and combined MiSTi burden and variance components using Fisher's Combination method. MiSTi models were adjusted for age at the reference time, sex, study, genotyping phase, total energy consumption, and the principal components to account for potential population substructure. The Bonferroni corrected threshold of P = 1.03 x 10^{-5} is labeled in red.

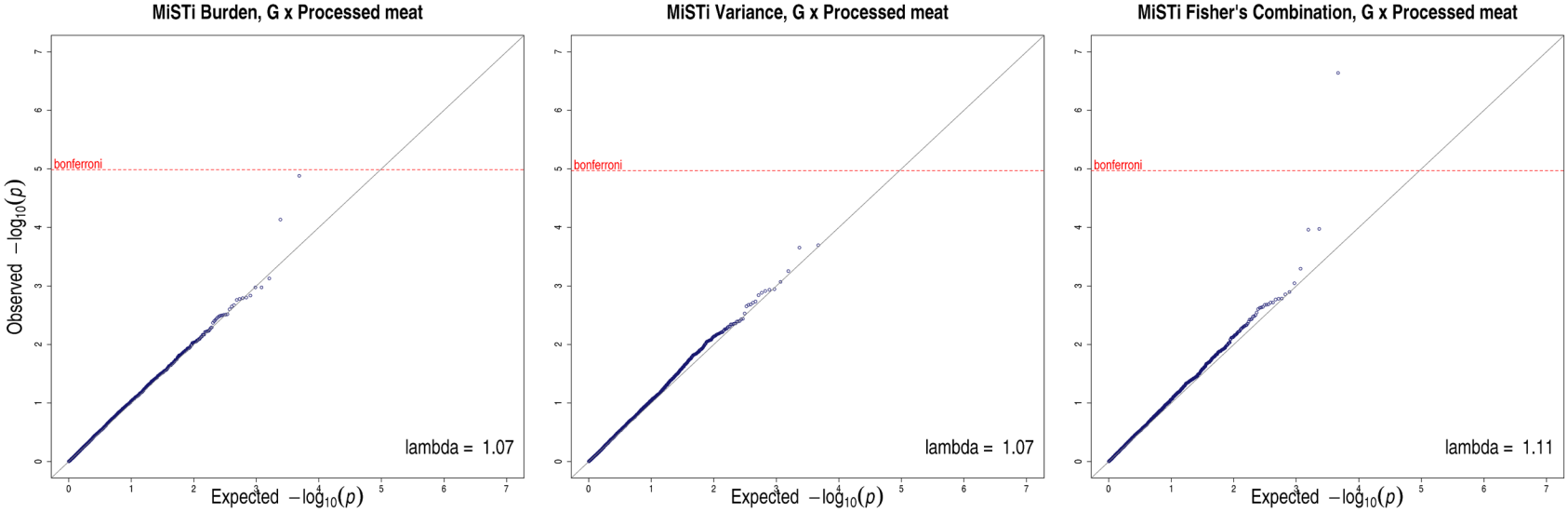


Table 2.2 MiSTi results for the Rac Family Small GTPase 1 (*RAC1*) gene-set and sex- and study- specific quartiles of processed meat intake.

RAC1 locus	*RAC1* gene-set variants	PrediXcan *RAC1* R^2	Burden P value	Variance P value	Fisher's Combined P value
7p22.1	N = 26	0.133	7.37×10^{-5}	2.01×10^{-4}	2.30×10^{-7}

Figure 2.8 Q-Q plots from gene-set interaction tests (n=4840) with sex- and -study specific quartiles of vegetable intake for risk of CRC. Results are from the MiSTi burden component, MiSTi variance component, and combined MiSTi burden and variance components using Fisher's Combination method. MiSTi models were adjusted for age at the reference time, sex, study, genotyping phase, total energy consumption, and the principal components to account for potential population substructure. The Bonferroni corrected threshold of $P = 1.03 \times 10^{-5}$ is labeled in red.

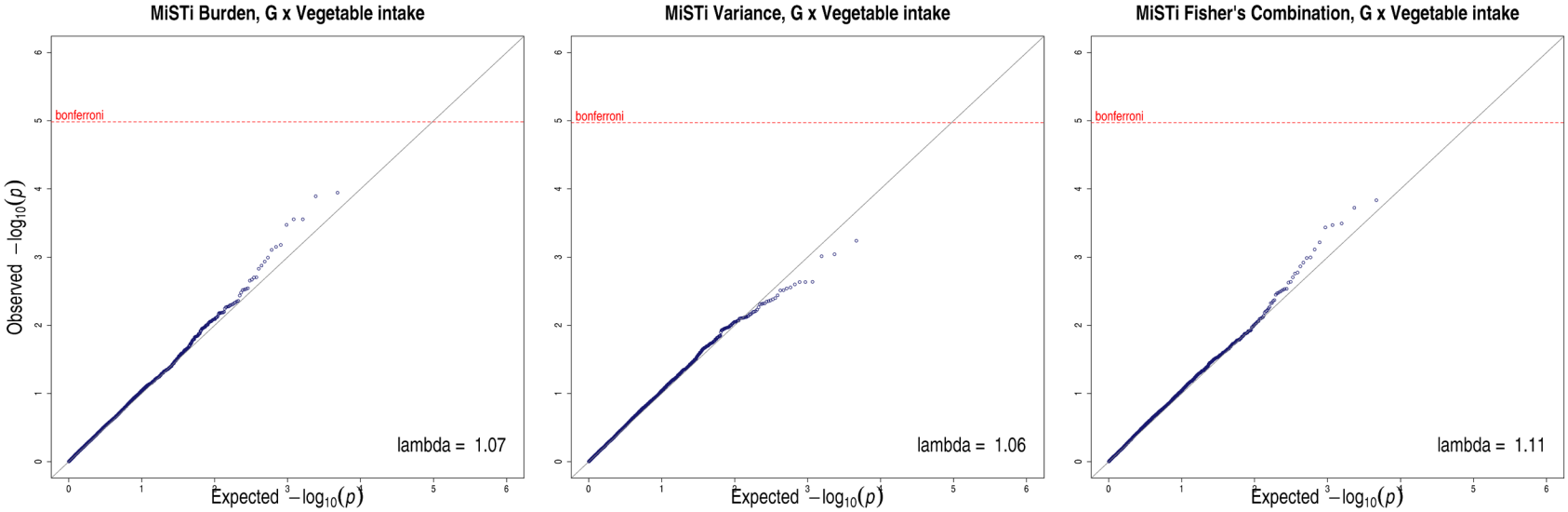


Figure 2.9 Q-Q plots from gene-set interaction tests (n=4840) with sex- and -study specific quartiles of fruit intake for risk of CRC. Results are from the MiSTi burden component, MiSTi variance component, and combined MiSTi burden and variance components using Fisher's Combination method. MiSTi models were adjusted for age at the reference time, sex, study, genotyping phase, total energy consumption, and the principal components to account for potential population substructure. The Bonferroni corrected threshold of $P = 1.03 \times 10^{-5}$ is labeled in red.

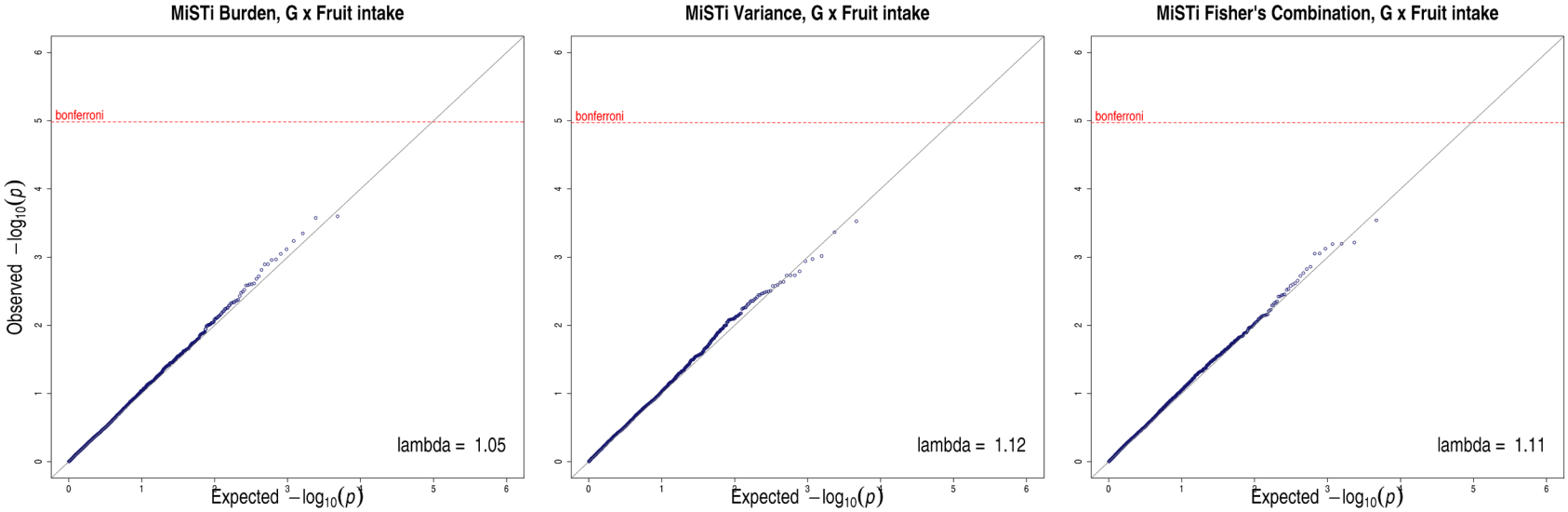


Figure 2.10 Q-Q plots from gene-set interaction tests (n=4840) with sex- and -study specific quartiles of dietary fiber intake for risk of CRC. Results are from the MiSTi burden component, MiSTi variance component, and combined MiSTi burden and variance components using Fisher's Combination method. MiSTi models were adjusted for age at the reference time, sex, study, genotyping phase, total energy consumption, and the principal components to account for potential population substructure. The Bonferroni corrected threshold of $P = 1.03 \times 10^{-5}$ is labeled in red.

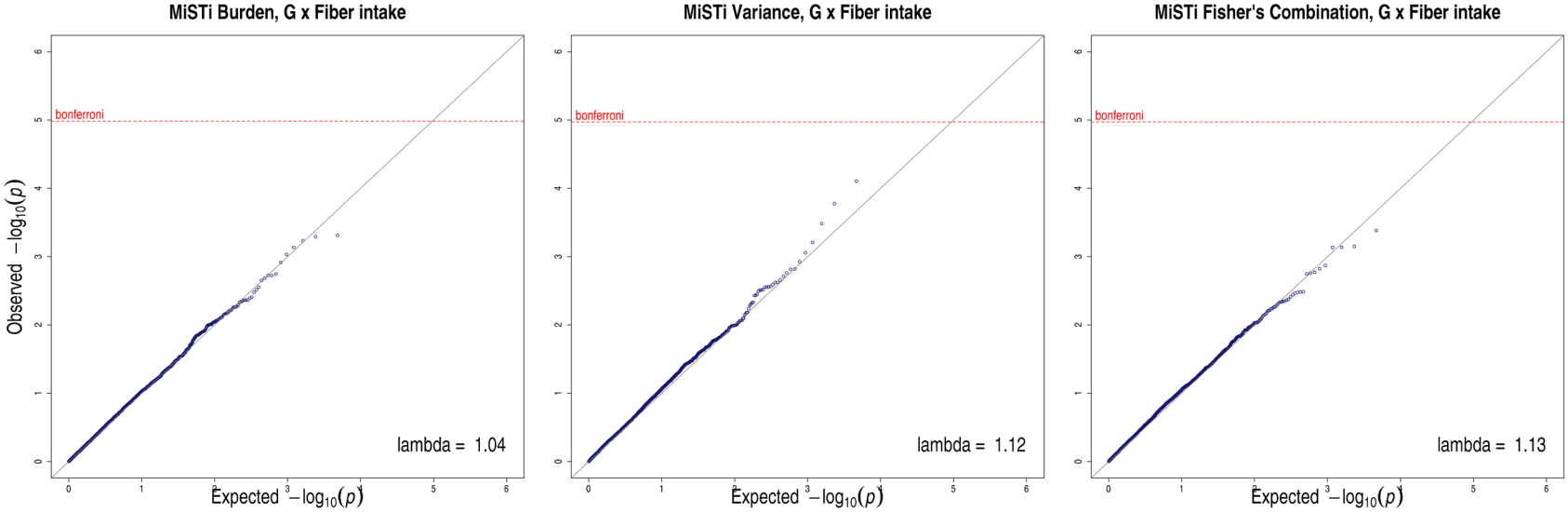


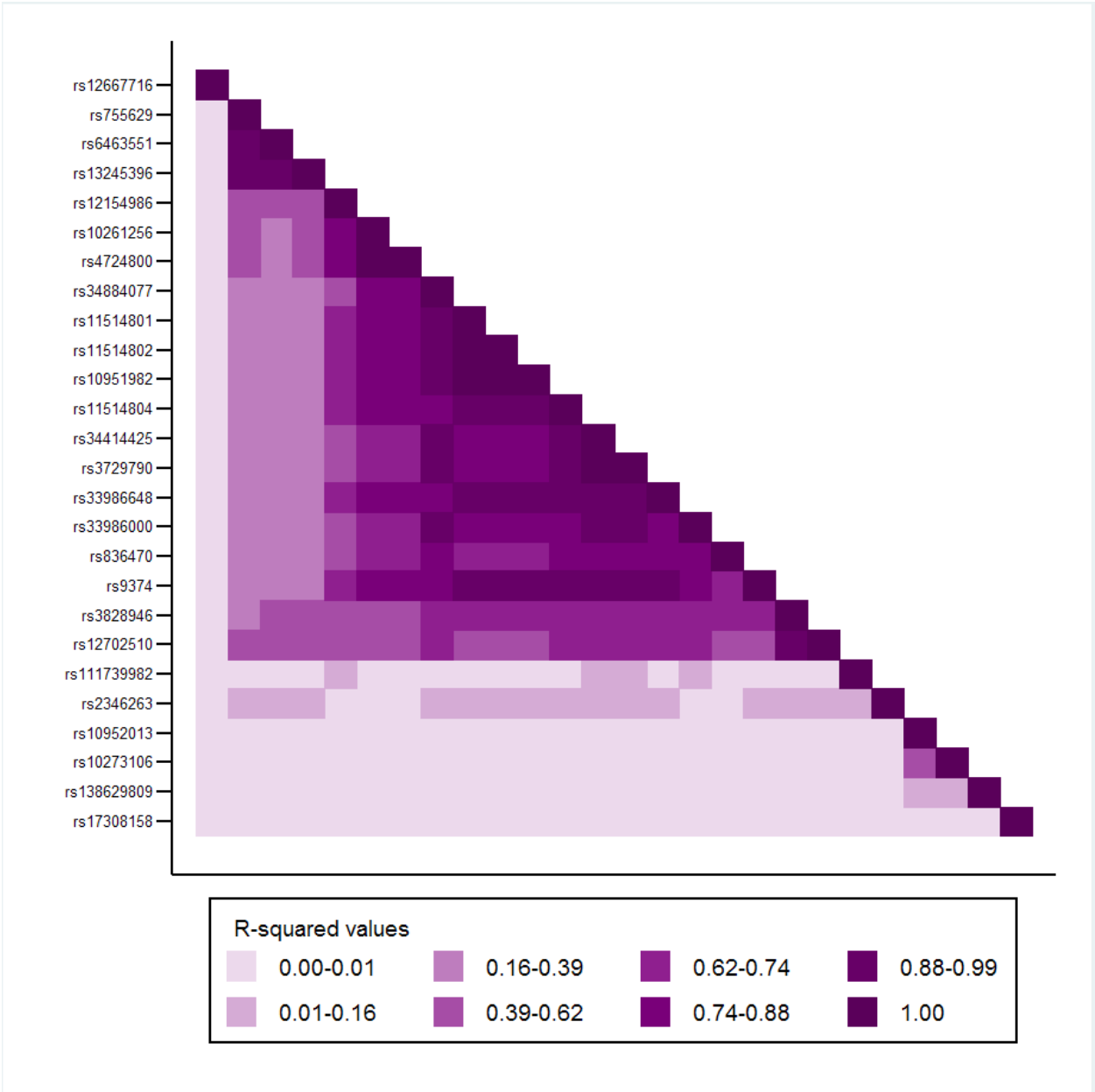
Table 2.3 Generalized linear model regression results for *RAC1* interaction effects with processed meat intake.

***RAC1* - Rac Family Small GTPase 1**								
β^E =0.206								
Variant	**Weight**	**Effect Allele Frequency**	**β^G**	**SE^G**	**P value**	**β^{GxE}**	**SE^{GxE}**	**P value**
rs10261256	-0.034	0.22	0.562	4.586	9.02×10^{-1}	-0.36	4.803	9.40×10^{-1}
rs10951982	-0.011	0.21	-7.444	14.03	5.96×10^{-1}	4.913	14.054	7.27×10^{-1}
rs10952013	0.024	0.35	0.052	0.038	1.76×10^{-1}	-0.063	0.035	7.29×10^{-02}
rs11514801	-0.011	0.21	-22.539	43.394	6.03×10^{-1}	-2.351	42.369	9.56×10^{-1}
rs11514802	-0.013	0.21	29.347	54.612	5.91×10^{-1}	-1.958	52.83	9.70×10^{-1}
rs11514804	-0.006	0.21	-0.030	1.258	9.81×10^{-1}	-0.279	1.417	8.44×10^{-1}
rs12154986	-0.013	0.23	-0.014	0.077	8.61×10^{-1}	0.044	0.08	5.83×10^{-1}
rs12667716	0.021	0.13	0.041	0.032	2.03×10^{-1}	-0.026	0.033	4.26×10^{-1}
rs12702510	0.040	0.73	-0.079	0.094	4.00×10^{-1}	-0.016	0.124	9.00×10^{-1}
rs13245396	-0.019	0.34	-0.403	2.201	8.55×10^{-1}	1.170	2.406	6.27×10^{-1}
rs138629809	0.078	0.04	-0.006	0.031	8.45×10^{-1}	0.020	0.031	5.28×10^{-1}
rs17308158	-0.005	0.39	0.000	0.031	9.97×10^{-1}	0.002	0.035	9.62×10^{-1}
rs2346263	0.012	0.29	0.125	0.032	8.15×10^{-5}	-0.127	0.033	1.32×10^{-04}
rs33986000	-0.011	0.19	-0.253	0.184	1.69×10^{-1}	0.145	0.193	4.53×10^{-1}
rs33986648	-0.007	0.21	0.516	1.215	6.71×10^{-1}	-0.237	1.377	8.63×10^{-1}
rs34414425	-0.004	0.20	-1.454	3.636	6.89×10^{-1}	-0.382	3.942	9.23×10^{-1}
rs34884077	-0.011	0.20	0.203	0.606	7.37×10^{-1}	-0.173	0.644	7.88×10^{-1}
rs3729790	-0.001	0.20	1.510	3.594	6.74×10^{-1}	0.402	3.911	9.18×10^{-1}
rs3828946	-0.069	0.24	0.053	0.107	6.19×10^{-1}	-0.124	0.114	2.77×10^{-1}
rs4724800	-0.084	0.22	-0.330	4.576	9.42×10^{-1}	0.189	4.792	9.69×10^{-1}
rs6463551	-0.010	0.34	0.497	2.222	8.23×10^{-1}	-1.307	2.432	5.91×10^{-1}
rs755629	-0.018	0.35	-0.108	0.188	5.67×10^{-1}	0.128	0.205	5.32×10^{-1}
rs836470	-0.016	0.24	0.015	0.078	8.49×10^{-1}	0.009	0.083	9.14×10^{-1}
rs9374	-0.072	0.20	-0.082	0.497	8.69×10^{-1}	0.046	0.539	9.32×10^{-1}

Estimates for genetic main effects (β^G) and interaction effects (β^{GxE}) for each variant from a generalized linear model adjusted for age at the referent time, sex, study, genotyping phase, total energy consumption, and principal components to account for potential population substructure.

Figure 2.11 LD heat map for variants in *RAC1* gene-set

Using R^2 values as a measure of linkage disequilibrium (LD), this heatmap shows a matrix of the strength of correlation between variants within the *RAC1* variant set. R^2 values were calculated based on variant allele frequencies in the European (EUR) sub-population of the 1000 Genomes Project.



REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394-424.

2. Peters U, Bien S, Zubair N. Genetic architecture of colorectal cancer. *Gut.* 2015;64(10):1623-1636.

3. Song M, Garrett WS, Chan AT. Nutrients, foods, and colorectal cancer prevention. *Gastroenterology.* 2015;148(6):1244-1260 e1216.

4. Randi G, Edefonti V, Ferraroni M, La Vecchia C, Decarli A. Dietary patterns and the risk of colorectal cancer and adenomas. *Nutr Rev.* 2010;68(7):389-408.

5. Bouvard V, Loomis D, Guyton KZ, et al. Carcinogenicity of consumption of red and processed meat. *Lancet Oncol.* 2015;16(16):1599-1600.

6. Huxley RR, Ansary-Moghaddam A, Clifton P, Czernichow S, Parr CL, Woodward M. The impact of dietary and lifestyle risk factors on risk of colorectal cancer: a quantitative overview of the epidemiological evidence. *Int J Cancer.* 2009;125(1):171-180.

7. Potter JD. Vegetables, fruit, and cancer. *Lancet.* 2005;366(9485):527-530.

8. Manolio TA, Collins FS. Genes, environment, health, and disease: facing up to complexity. *Hum Hered.* 2007;63(2):63-66.

9. Al-Tassan NA, Whiffin N, Hosking FJ, et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci Rep.* 2015;5:10442.

10. Peters U, Jiao S, Schumacher FR, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology.* 2013;144(4):799-807 e724.

11. Real LM, Ruiz A, Gayan J, et al. A colorectal cancer susceptibility new variant at 4q26 in the Spanish population identified by genome-wide association analysis. *PLoS One.* 2014;9(6):e101178.

12. Schmit SL, Edlund CK, Schumacher FR, et al. Novel Common Genetic Susceptibility Loci for Colorectal Cancer. *J Natl Cancer Inst.* 2019;111(2):146-157.

13. Schmit SL, Schumacher FR, Edlund CK, et al. A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. *Carcinogenesis.* 2014;35(11):2512-2519.

14. Schumacher FR, Schmit SL, Jiao S, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun.* 2015;6:7138.

15. Wang M, Gu D, Du M, et al. Common genetic variation in ETV6 is associated with colorectal cancer susceptibility. *Nat Commun.* 2016;7:11478.

16. Zeng C, Matsuda K, Jia WH, et al. Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. *Gastroenterology.* 2016;150(7):1633-1645.

17. Sengupta S, Muir JG, Gibson PR. Does butyrate protect from colorectal cancer? *J Gastroenterol Hepatol.* 2006;21(1 Pt 2):209-218.

18. 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.

19. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med.* 2000;343(2):78-85.

20. Wang H, Burnett T, Kono S, et al. Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat Commun.* 2014;5:4613.

21. Figueiredo JC, Hsu L, Hutter CM, et al. Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet.* 2014;10(4):e1004228.

22. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med.* 2002;21(1):35-50.

23. Su YR, Di CZ, Hsu L, Genetics, Epidemiology of Colorectal Cancer C. A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics.* 2017;18(1):119-131.

24. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank.* 2015;13(5):307-308.

25. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science.* 2015;348(6235):648-660.

26. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. *Am J Hum Genet.* 2013;93(4):687-696.

27. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48(10):1284-1287.

28. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* 2016;48(10):1279-1283.

29. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.

30. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet.* 2015;47(9):1091-1098.

31. The Genomes Project C, Auton A, Abecasis GR, et al. A global reference for human genetic variation. *Nature.* 2015;526:68.

32. Ananthakrishnan AN, Du M, Berndt SI, et al. Red meat intake, NAT2, and risk of colorectal cancer: a pooled analysis of 11 studies. *Cancer Epidemiol Biomarkers Prev.* 2015;24(1):198-205.

33. Hutter CM, Liu Y, Duggan DJ, et al. Examination of Gene-Environment Interactions for Known Colorectal Cancer Susceptibility Loci. *American journal of epidemiology.* 2010;171:S65-S65.

34. Kantor ED, Hutter CM, Minnier J, et al. Gene-environment interaction involving recently identified colorectal cancer susceptibility Loci. *Cancer Epidemiol Biomarkers Prev.* 2014;23(9):1824-1833.

35. Koushik A, Hunter DJ, Spiegelman D, et al. Fruits, vegetables, and colon cancer risk in a pooled analysis of 14 cohort studies. *J Natl Cancer Inst.* 2007;99(19):1471-1483.

36. Park Y, Subar AF, Kipnis V, et al. Fruit and vegetable intakes and risk of colorectal cancer in the NIH-AARP diet and health study. *Am J Epidemiol.* 2007;166(2):170-180.

37. Team RC. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing;2013.

38. Fisher RA GS, Britain G, Généticien S. *Statistical methods for research workers.* Vol 14. Edinburgh: Oliver and Boyd; 1970.

39. Vieira AR, Abar L, Chan DSM, et al. Foods and beverages and colorectal cancer risk: a systematic review and meta-analysis of cohort studies, an update of the evidence of the WCRF-AICR Continuous Update Project. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO.* 2017;28(8):1788-1802.

40. Butler LM, Sinha R, Millikan RC, et al. Heterocyclic amines, meat intake, and association with colon cancer in a population-based study. *Am J Epidemiol.* 2003;157(5):434-445.

41. El-Bayoumy K, Sinha R. Molecular chemoprevention by selenium: a genomic approach. *Mutat Res.* 2005;591(1-2):224-236.

42. Joosen AM, Kuhnle GG, Aspinall SM, et al. Effect of processed and red meat on endogenous nitrosation and DNA damage. *Carcinogenesis.* 2009;30(8):1402-1407.

43. Bastide NM, Pierre FH, Corpet DE. Heme iron from meat and risk of colorectal cancer: a meta-analysis and a review of the mechanisms involved. *Cancer prevention research.* 2011;4(2):177-184.

44. Alisson-Silva F, Kawanishi K, Varki A. Human risk of diseases associated with red meat intake: Analysis of current theories and proposed role for metabolic incorporation of a non-human sialic acid. *Mol Aspects Med.* 2016;51:16-30.

45. Tricker AR. N-nitroso compounds and man: sources of exposure, endogenous formation and occurrence in body fluids. *Eur J Cancer Prev.* 1997;6(3):226-268.

46. Schwartz S, Ellefson M. Quantitative fecal recovery of ingested hemoglobin-heme in blood: comparisons by HemoQuant assay with ingested meat and fish. *Gastroenterology.* 1985;89(1):19-26.

47. Young GP, Rose IS, St John DJ. Haem in the gut. I. Fate of haemoproteins and the absorption of haem. *J Gastroenterol Hepatol.* 1989;4(6):537-545.

48. Santarelli RL, Pierre F, Corpet DE. Processed meat and colorectal cancer: a review of epidemiologic and experimental evidence. *Nutr Cancer.* 2008;60(2):131-144.

49. Xue X, Shah YM. Intestinal iron homeostasis and colon tumorigenesis. *Nutrients.* 2013;5(7):2333-2351.

50. Lewin MH, Bailey N, Bandaletova T, et al. Red meat enhances the colonic formation of the DNA adduct O6-carboxymethyl guanine: implications for colorectal cancer risk. *Cancer Res.* 2006;66(3):1859-1865.

51. Gilsing AM, Fransen F, de Kok TM, et al. Dietary heme iron and the risk of colorectal cancer with specific mutations in KRAS and APC. *Carcinogenesis.* 2013;34(12):2757-2766.

52. Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Annals of internal medicine.* 2011;154(1):22-30.

53. Leufkens AM, van Duijnhoven FJ, Woudt SH, et al. Biomarkers of oxidative stress and risk of developing colorectal cancer: a cohort-nested case-control study in the European Prospective Investigation Into Cancer and Nutrition. *Am J Epidemiol.* 2012;175(7):653-663.

54. Martin OCB, Olier M, Ellero-Simatos S, et al. Haem iron reshapes colonic luminal environment: impact on mucosal homeostasis and microbiome through aldehyde formation. *Microbiome.* 2019;7(1):72.

55. Seinen ML, van Nieuw Amerongen GP, de Boer NK, van Bodegraven AA. Rac Attack: Modulation of the Small GTPase Rac in Inflammatory Bowel Disease and Thiopurine Therapy. *Mol Diagn Ther.* 2016;20(6):551-557.

56. Frenkel S, Bernstein CN, Sargent M, et al. Genome-wide analysis identifies rare copy number variations associated with inflammatory bowel disease. *PLoS One.* 2019;14(6):e0217846.

57. Ellenbroek SI, Collard JG. Rho GTPases: functions and association with cancer. *Clin Exp Metastasis.* 2007;24(8):657-672.

58. Fiegen D, Haeusler LC, Blumenstein L, et al. Alternative splicing of Rac1 generates Rac1b, a self-activating GTPase. *J Biol Chem.* 2004;279(6):4743-4749.

59. Orlichenko L, Geyer R, Yanagisawa M, et al. The 19-amino acid insertion in the tumor-associated splice isoform Rac1b confers specific binding to p120 catenin. *J Biol Chem.* 2010;285(25):19153-19161.

60. Matos P, Jordan P. Increased Rac1b expression sustains colorectal tumor cell survival. *Mol Cancer Res.* 2008;6(7):1178-1184.

61. Nie F, Zhao SY, Song FX, Li PW. Changes of cytoskeleton and cell cycle in Lovo cells via deletion of Rac1. *Cancer Biomark.* 2014;14(5):335-342.

62. Bid HK, Roberts RD, Manchanda PK, Houghton PJ. RAC1: an emerging therapeutic option for targeting cancer angiogenesis and metastasis. *Mol Cancer Ther.* 2013;12(10):1925-1934.

63. Myant KB, Cammareri P, McGhee EJ, et al. ROS production and NF-kappaB activation triggered by RAC1 facilitate WNT-driven intestinal stem cell proliferation and colorectal cancer initiation. *Cell Stem Cell.* 2013;12(6):761-773.

64. Espina C, Cespedes MV, Garcia-Cabezas MA, et al. A critical role for Rac1 in tumor progression of human colorectal adenocarcinoma cells. *Am J Pathol.* 2008;172(1):156-166.

65. Gehart H, Clevers H. Tales from the crypt: new insights into intestinal stem cells. *Nat Rev Gastroenterol Hepatol.* 2019;16(1):19-34.

66. Abuli A, Castells A, Bujanda L, et al. Genetic Variants Associated with Colorectal Adenoma Susceptibility. *PLoS One.* 2016;11(4):e0153084.

67. Zhang Y, Yan L, Zeng J, et al. Pan-cancer analysis of clinical relevance of alternative splicing events in 31 human cancers. *Oncogene.* 2019;38(40):6678-6695.

68. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. *Ann Epidemiol.* 1994;4(1):1-10.

69. Alpha-Tocopherol BCCPSG. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med.* 1994;330(15):1029-1035.

70. Calle EE, Rodriguez C, Jacobs EJ, et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer.* 2002;94(9):2490-2501.

71. Newcomb PA, Baron J, Cotterchio M, et al. Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev.* 2007;16(11):2331-2343.

72. Figueiredo JC, Lewinger JP, Song C, et al. Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. *Cancer Epidemiol Biomarkers Prev.* 2011;20(5):758-766.

73. Lilla C, Verla-Tebit E, Risch A, et al. Effect of NAT1 and NAT2 genetic polymorphisms on colorectal cancer risk associated with exposure to tobacco smoke and meat consumption. *Cancer Epidemiol Biomarkers Prev.* 2006;15(1):99-107.

74. Slattery ML, Berry TD, Potter J, Caan B. Diet diversity, diet composition, and risk of colon cancer (United States). *Cancer Causes Control.* 1997;8(6):872-882.

75. Kury S, Buecher B, Robiou-du-Pont S, et al. Combinations of cytochrome P450 gene polymorphisms enhancing the risk for sporadic colorectal cancer related to red meat consumption. *Cancer Epidemiol Biomarkers Prev.* 2007;16(7):1460-1467.

76. Le Marchand L, Hankin JH, Wilkens LR, et al. Combined effects of well-done red meat, smoking, and rapid N-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev.* 2001;10(12):1259-1266.

77. Rimm EB, Stampfer MJ, Colditz GA, Chute CG, Litin LB, Willett WC. Validity of self-reported waist and hip circumferences in men and women. *Epidemiology.* 1990;1(6):466-473.

78. Li D, Conti DV. Detecting gene-environment interactions using a combined case-only and case-control approach. *Am J Epidemiol.* 2009;169(4):497-504.

79. Helmus DS, Thompson CL, Zelenskiy S, Tucker TC, Li L. Red meat-derived heterocyclic amines increase risk of colon cancer: a population-based case-control study. *Nutr Cancer.* 2013;65(8):1141-1150.

80. Milne RL, Fletcher AS, MacInnis RJ, et al. Cohort Profile: The Melbourne Collaborative Cohort Study (Health 2020). *Int J Epidemiol.* 2017;46(6):1757-1757i.

81. Giles GG, English DR. The Melbourne Collaborative Cohort Study. *IARC Sci Publ.* 2002;156:69-70.

82. Poynter JN, Gruber SB, Higgins PD, et al. Statins and the risk of colorectal cancer. *N Engl J Med.* 2005;352(21):2184-2192.

83. Kolonel LN, Henderson BE, Hankin JH, et al. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol.* 2000;151(4):346-357.

84. Woods MO, Younghusband HB, Parfrey PS, et al. The genetic basis of colorectal cancer

in a population-based incident cohort with a high rate of familial disease. *Gut.* 2010;59(10):1369-1377.

85. Belanger CF, Hennekens CH, Rosner B, Speizer FE. The nurses' health study. *Am J Nurs.* 1978;78(6):1039-1040.

86. Cotterchio M, McKeown-Eyssen G, Sutherland H, et al. Ontario familial colon cancer registry: methods and first-year response rates. *Chronic Dis Can.* 2000;21(2):81-86.

87. Christen WG, Gaziano JM, Hennekens CH. Design of Physicians' Health Study II--a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Ann Epidemiol.* 2000;10(2):125-134.

88. Hennekens CH, Eberlein K. A randomized trial of aspirin and beta-carotene among U.S. physicians. *Prev Med.* 1985;14(2):165-168.

89. Newcomb PA, Zheng Y, Chia VM, et al. Estrogen plus progestin use, microsatellite instability, and the risk of colorectal cancer in women. *Cancer Res.* 2007;67(15):7534-7539.

90. Gohagan JK, Prorok PC, Hayes RB, Kramer BS, Prostate LC, Ovarian Cancer Screening Trial Project T. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control Clin Trials.* 2000;21(6 Suppl):251S-272S.

91. Prorok PC, Andriole GL, Bresalier RS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin Trials.* 2000;21(6 Suppl):273S-309S.

92. Landi MT, Chatterjee N, Yu K, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet.* 2009;85(5):679-691.

93. Harris H, Håkansson N, Olofsson C, et al. The Swedish mammography cohort and the cohort of Swedish men: study design and characteristics of two populationbased longitudinal cohorts. *OA Epidemiology.* 2013;1(2):16.

94. White E, Patterson RE, Kristal AR, et al. VITamins And Lifestyle cohort study: study design and characteristics of supplement users. *Am J Epidemiol.* 2004;159(1):83-93.

95. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials.* 1998;19(1):61-109.

96. Hays J, Hunt JR, Hubbell FA, et al. The Women's Health Initiative recruitment methods and results. *Ann Epidemiol.* 2003;13(9 Suppl):S18-77.

97. Bergstralh EH KJ. Computerized matching of cases to controls. *Technical Report Number 56.* 1995.

CHAPTER 3. INTERACTIONS BETWEEN PREDICTED GENE EXPRESSION WITH CIRCULATING C-REACTIVE PROTEIN CONCENTRATION AND RISK OF COLORECTAL CANCER

3.1 ABSTRACT

Background: There is robust evidence that chronic inflammation has a role in colorectal cancer (CRC) risk. C-reactive protein (CRP) is a marker for chronic inflammation and pre-diagnostic concentrations have been associated with increased risk of CRC in some but not all studies, suggesting that interactions with genetic variants may account for part of this heterogeneity.

Methods: We harmonized pre-diagnostic circulating CRP data from five studies within the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) including 2,625 colorectal tumors, and 2,798 controls of European ancestry. We estimated genetically determined gene expression levels using PrediXcan based upon transcriptome data of colon tissues from the Genotype-Tissue Expression (GTEx) Project for all genes with heritability ≥1%. Using the Mixed Score Test for Interactions, we tested if genetically predicted gene expression modified the effect of circulating CRP on risk of CRC. We tested 4839 genes. We set a False Discovery Rate (FDR) threshold to 0.20 to account for multiple testing.

Results: In multivariable adjusted fixed-effect meta-analysis, a two-fold-higher circulating CRP concentration was associated with a moderately higher risk of colorectal cancer [OR:1.04 (95% CI:1.00, 1.08)]. There was significant heterogeneity in study-specific effect sizes for circulating CRP concentrations [$P = 7.6 \times 10^{-3}$]. We detected no significant interactions between circulating CRP concentrations and variants predictive of gene expression in the colon at an FDR < 0.20.

Conclusions: Circulating CRP concentrations were modestly associated with CRC risk in our study, but we did not find evidence that genetically determined gene expression in the colon

influences the association between CRP and CRC risk. Circulating CRP concentration may not correlate well with the inflammation processes most etiologically relevant to colorectal tumorigenesis.

3.2 Introduction

Colorectal cancer (CRC) is the second most commonly diagnosed cancer worldwide when considering men and women combined, as well as the second leading cause of cancer deaths[1]. Colorectal tumorigenesis is a multi-factorial process with both genetic (G) and environmental factors (E) contributing to disease development[2]. A relationship between chronic inflammation and cancer is supported by evidence from both epidemiological and experimental studies, yet the mechanisms that underlie the association between inflammation and CRC remain uncertain[3-5]. Chronic inflammation as a risk factor for CRC is supported by studies of individuals with inflammatory bowel disease (IBD), which have shown that IBD patients are at higher risk of CRC, with an increase in risk that is proportional to the duration and anatomical extent of their disease[6]. The use of non-steroidal anti-inflammatory drugs (NSAIDs) is associated with decreased risk of CRC[7-12]. NSAIDs inhibit the production of prostanoids associated with inflammation and tumor growth that are overexpressed in ~50% of colorectal adenomas and 85% of carcinomas[13-15]. Moreover, metabolic diseases characterized by a state of chronic low-grade inflammation[16] such as insulin resistance and obesity[17] are associated with increased CRC risk.

C-reactive protein (CRP) is an acute-phase reactant protein produced in the liver as a response to infection, tissue injury, and systemic inflammation. CRP is the biomarker frequently used to monitor chronic inflammation[18] and circulating CRP concentrations have been associated with an increased risk of CRC in some but not all studies[19-21]. Interactions with genetic variants may contribute to this heterogeneity in the association between CRP and CRC risk. Genome-wide association studies (GWAS) of sporadic CRC[22-30] have discovered more than 150 heritable loci that predict only a fraction of the estimated 35% heritability of CRC risk[31,32]. For a multi-factorial disease like CRC, gene-by-environment (GxE) interaction studies can offer important

insight to CRC etiology and risk through the identification of novel CRC risk loci. CRC risk loci may suggest pathways that influence CRC development to inform both prevention and treatment of CRC[33].

In this GxE interaction study of CRP concentrations and CRC risk in 2,625 CRC cases and 2,798 controls from five studies within the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), we tested whether CRP (as a proxy measure for chronic inflammation), interacts with genetic variation to influence CRC risk. To increase statistical power, we used a novel mixed effects, set-based approach that incorporated *a priori* functional information for variants predictive of gene expression in colon tissue based on data from the GTEx Project[34].

3.3 METHODS

Circulating CRP Data

Our study sample was comprised of 2,583 CRC cases, 42 adenoma cases, and 2,798 controls all of European ancestry. Individual level CRP measurements were collected from each study for pooled analysis following an iterative process established with our previous data harmonization of demographics, clinical data and lifestyle risk factors. Contributing studies had CRP measurements from high-sensitivity assay methods, with intra-assay coefficients of variance ranging from 0.9% to 7.8% (Table 3.2). To maximize the sample size from WHI, we selected participants across several previously published WHI CRP studies (Supplemental Table 4). We selected a single CRP baseline measurement from each WHI participant (Supplemental Figure 2). All CRP data were written to a common data platform, transformed, and combined into a single dataset. Data were then reviewed once more for quality assurance, using range and logic checks as well as an examination of intra- and inter-study variable distributions by outcome, sex,

and other potential confounders. To exclude gross outliers, CRP values were truncated to a maximum value of 100 mg/L. CRP values were log_2-transformed because the distribution of CRP values are known to be positively skewed and this transformation rescales the CRP values to be more consistent with a normal distribution[35]. The log_2-transformed CRP data were checked for normal distribution by plotting frequency histograms by study and outcome, prior to subsequent statistical analysis.

Genotype Data

Genotype data was generated from germline DNA on several Illumina platforms (300K, 550K, 550K duo, 610K, OmniExpress, OmniExpressExome, and Oncoarray+custom iSelect). Sample collection, genotyping, and quality control (QC) including average sample and variant call rates, and concordance rates for blinded duplicates have been previously described in detail for all GWAS data[2,24,29,30]. To summarize, we excluded duplicates, close relatives defined as individuals that are second degree or more closely related, and samples with discrepancies between reported and genotypic sex, and samples with call rate <98%. Directly genotyped variants were excluded for a call rate <98%, Hardy-Weinberg Equilibrium (HWE) with $P < 1 \times 10^{-4}$ in controls, and an effect allele count less than 3. After QC, genotype data was imputed to the Haplotype Reference Consortium (HRC version r1.1) using the Michigan Imputation Server with phasing option set to ShapeIT v2.r790[36-38]. Variants were restricted to imputation accuracy R2 > 0.3. We performed principal components analysis (PCA) using PLINK 2.0[39] and analyses were restricted to samples that clustered with European descent in PCA. Variants were annotated using the human assembly GRCh37 (h19).

Functional Information

Variant weights from the PrediXcan[40] approach for estimating genetically determined gene expression were used as *a priori* functional information for set-based GxE testing. We downloaded the tissue-specific gene expression models of the PrediXcan approach from the publicly accessible PredictDB Data Repository (http://hakyimlab.org/predictdb/). Gene expression prediction models for colon transverse tissues were built using data from the GTEx Project V6p[34] of predominantly European ancestry samples (n=169). The PrediXcan reference cohort, datasets, and model building have previously been described in detail[40]. Briefly, Gamazon et al. developed additive models of gene expression levels in PrediXcan using jointly measured genome-wide genotype data and RNA-seq data[40]. We excluded the colon sigmoid PrediXcan models from our analyses because sample attributes available from the GTEx Portal indicated that the sampling target for the sigmoid colon was muscularis and not mucosa, which is the tissue of interest for CRC (GTEx_Data_V6_Annotations_SampleAttributesDS.txt). Samples for the GTEx transverse colon reference data included the entire colonic wall, including mucosa. We selected the transverse colon (hereafter referred to as "colon" in this paper) prediction models trained with genotype data imputed to the 1000 Genome Project v3[41]. The models were comprised of 4,878 gene sets with 145,258 unique variants. Of those, 99.2% of genes and 86.9% of variants were covered in our genotype data. Variants in the PrediXcan models that did not intersect with our genotype data were excluded from analyses.

Epidemiological Data

Basic demographics, lifestyle and dietary risk factors were collected through self-report using in-person interviews and/or structured questionnaires. Harmonization of individual level data was performed at the GECCO data coordinating center. Data harmonization comprised a multi-step process for data reconciliation of each study's unique protocols and data-collection instruments

(Supplemental Figure 1). All CRC cases were defined as colorectal adenocarcinoma and confirmed by medical records, pathologic reports, or death certificate. Advanced adenomas were included because they are clinically relevant precursors to CRC, and were confirmed by medical records, histopathology, or pathologic reports. Controls for advanced adenoma cases had a negative colonoscopy except for controls matched to cases with distal adenoma, which either had a negative sigmoidoscopy or colonoscopy exam. All studies received ethical approval by their respective Institutional Review Boards. Studies are summarized in detail in (Supplemental Text 1). All variables for analysis were collected at study referent time, which was defined as study enrollment or blood collection. Age at referent time for the Hawaiian Case Control Adenoma Study (Hawaii CCS) was defined as age at diagnosis for adenoma cases or age at selection for adenoma controls. Time from blood draw for CRP assay to CRC diagnosis for European Prospective Investigation into Cancer and Nutrition (EPIC), Health Professionals Follow-up Study (HPFS), Nurses' Health Study (NHS), and Women's Health Initiative (WHI) was measured in years. In addition to age and sex (male/female), an *a priori* list of potential confounders was examined for candidate adjustment variables in modeling the association between CRP and CRC. These variables included: BMI (kg/m^2), exercise (sedentary/non-sedentary), diabetes (yes/no), smoking status (never/former/current), alcohol use (non-drinker, 1-28g/day,>28g/day), regular aspirin or NSAID use (yes/no), diabetes (yes/no), and postmenopausal hormone (PMH) use in women. The exercise variable was defined as any physical activity ≥ 1 hour per week. Missing BMI was replaced by multivariate imputation by chained equations (MICE) using the R mice package[42]. Imputation was done ten times according to the dependence of BMI on age, sex, study, height (cm), and total energy (kcal/day).

Statistical Analyses

Statistical analyses of all data were conducted centrally at the GECCO coordinating center on individual-level data to ensure a consistent analytical approach. All analyses were conducted using R[43]. Multivariable logistic regression was used to estimate study-specific odds ratios (ORs) and 95% confidence intervals (CIs) for the association between log_2-transformed circulating CRP and CRC risk; study-specific estimates were combined using fixed-effects meta-analysis. A minimally adjusted model for logistic regression included variables for age at the referent time, sex, and study. A change in estimate approach was then used to select additional adjustment variables from BMI, exercise, diabetes, smoking status, alcohol use, regular aspirin/NSAID use, smoking status, diabetes, and PMH use in women. The fully adjusted logistic regression model included age, sex, study, BMI, and PMH use. We used the set-based approach Mixed Effects Score Test for Interactions (MiSTi) to test for gene-CRP interactions[44]. MiSTi provides a unified hierarchical regression framework for modeling GxE effects that has two components: interaction of E with genetically predicted gene expression as fixed effects (burden component), and residual heterogeneous GxE effects as random effects (variance component). We used the colon PrediXcan variant weights to calculate the genetically predicted gene expression. We assessed the gene-aggregated variant interaction effect with circulating CRP on CRC risk. Genotypes were treated as continuous variables (i.e. log-additive effects). We used the expected number of copies of the imputed variant effect allele (the "dosage"), which has been shown to give unbiased test statistics[45]. We adjusted the MiSTi model for age at the referent time, sex, study, BMI, PMH use, genotyping platform and principal components (PCs). We used Fisher's combination method[46] to calculate an overall p-value combining the p-value from both the fixed and random effects. We assessed the potential inflation due to unaccounted for confounders by using quantile-quantile (Q-Q) plots of the p-values for the fixed effects and random effects. We

also calculated genomic control (λ). Of 4878 gene-sets, 38 had a small number of variants (1 to 4) which created too sparse data cells for testing with MiSTi, and thus were excluded. We used a False Discovery Rate (FDR) of less than 0.20 as a threshold to determine significant gene-CRP interactions.

3.4 RESULTS

There were more women (72%) than men in our study overall. The mean age was similar between cases (63.8 years) and controls (63.4 years). The average BMI was higher among cases (27.5 kg/m^2) than controls (26.6 kg/m^2). A higher proportion of cases were sedentary (46.7%), and had diabetes (6.7%). The median time between blood sample collection and cancer diagnosis among EPIC, HPFS, NHS, and WHI was 5 years with an IQR of 3 to 8 years. Mean circulating CRP levels were higher among cases (4.15 mg/L) than controls (3.28 mg/L), and this was also true at the study-specific level except for the NHS. In a fixed effect meta-analysis (Figure 3.1), after adjustment for age, sex, study, BMI, and PMH use log2-transformed CRP levels were marginally associated with risk of CRC [OR:1.04 (95% CI:1.00, 1.08)]. We found that additional adjustment with smoking status, or alcohol use among studies with alcohol data available (EPIC, HPFS, NHS, and WHI) did not change the estimate of effect between CRP and CRC. Tests for heterogeneity in the meta-analyses indicated statistically significant variation in effect sizes in the fully adjusted model [$P = 7.6 \times 10^{-3}$]. The primary driver of heterogeneity in effect size of CRP was the NHS, which had an inverse association between CRP levels and CRC risk [OR:0.88 (95% CI:0.79, 0.97)]. To explore the heterogeneity, we performed a sensitivity analysis. After excluding the NHS from meta-analysis, the positive association between CRP and CRC was no longer heterogenous (P = 0.675) and the risk estimate for a two-fold increase in CRP concentration increased by 3% [OR:1.07 (95% CI:1.02, 1.12)].

In total, we tested 4,839 gene-based variant sets predictive of gene expression in the colon using MiSTi. The gene-sets had predictive performance (R^2) values that ranged from 0.018 to 0.699, with a median value of 0.062. The number of variants in each gene set ranged from 1 to 227 with a median of 25 variants. No statistically significant interactions at an FDR <0.20 were detected between genes and circulating CRP on risk of CRC when minimally adjusted for age at the referent time, sex, BMI, PMH, smoking, study, genotyping platform, total energy intake and PCs. Quantile-quantile (Q-Q) plots of p-values for the tests for the burden component, variance component, and combined burden and variance components using Fisher's combination method are presented in Figure 3.2. The Q-Q plots did not indicate premature departures of the distribution of p-values from null distribution, consistent with the genomic control values (λ range: 1.00 – 1.10). As with the meta-analysis of the association between CRP and CRC, we performed a sensitivity analysis with MiSTi for the interaction between gene-sets and CRP concentration on risk of CRC. The exclusion of NHS did not change results, and there were no detected interactions at an FDR <0.20 between genes and circulating CRP on risk of CRC (Figure 3.3). Genomic control values were slightly improved in the interaction tests excluding the NHS (λ range: 1.00 – 1.03).

3.5 Discussion

In our meta-analysis of CRP concentration and CRC risk in 5 epidemiologic studies, we observed a marginally statistically significant 4% increase in CRC risk per a two-fold increase in CRP concentration (mg/L). Previous studies of CRP and CRC risk have reported no association[47-50], as well as inverse associations in the Women's Health Study[51] and the NHS study of inflammatory markers[52]. Two previous studies of CRP and adenomas, including the

Hawaii CCS, did not find any association[53,54]. We observed statistically significant heterogeneity across studies primarily driven by NHS. Our results for the NHS was similar to a previous NHS CRP analysis in the NHS, where CRP was inversely associated with CRC[52]. Despite the inter-study heterogeneity in the association between CRP and CRC risk that we observed, our summary estimate for the association between CRP and CRC is similar in scale to previous studies and a meta-analysis that observed greater than 10% increases in risk with log-transformed CRP[19,53]. Our pooled GxE analysis did not detect any statistically significant interaction effects between CRP concentration and predicted gene expression in the colon on risk of CRC.

A limitation our analysis is the measurement error for CRP, which can come from laboratory, storage or batch effects. However, in our study the sample collection was largely done in a uniform manner across cases and controls within each study. Intra-assay coefficients of were overall low. The highest CV (7.8%) came from PLCO which represented approximately 10% of the overall sample (Table 3.2). Additionally, a single CRP measurement may not necessarily represent general systemic inflammation over time. As an indicator for chronic inflammation this could lead to an under or over measurement of "inflammation" because CRP concentrations can change in response to injury or from acute infection, increasing as much as 1000-fold from basal concentrations, and decrease back to baseline over several days[55]. Aside from infection or injury, however, CRP is expected to not have a great deal of intra-individual variation seasonally or diurnally among healthy people[56,57]. Timing for the CRP measurement in relationship to CRC diagnosis is also important because of the pro-inflammatory potential of tumors to elevate systemic measures of inflammation[58]. The average cancer sojourn time, the period when a cancer is screening detectable before it is clinically detectable, has been estimated to be at least 3 to 5

years[59]. An association between CRP and CRC in some studies could be due to reversed causality where increased CRP concentrations are a consequence of inflammation from precancerous lesions or pre-clinical disease. To maximize our study sample, we did not exclude subjects based on time between CRP measurement and CRC diagnosis. However, the median time between baseline CRP measurement and CRC diagnosis in our study was 5.7 years. We also included CRP measurements within 1 year of advanced adenoma diagnosis from the Hawaii CCS in our study. Advanced adenomas are important precursors to CRC, with transition rates over 10 years to CRC that are estimated to be between 25% and 43%, depending on age and sex[60]. Not all adenomas progress to cancer, but it is estimated that 85% of CRCs develop from advanced adenomas[60]. The average dwell time, the time period from adenoma initiation to development of clinical cancer is thought to be greater than 10 years, perhaps as much as 25 years[60,61]. Because of the long latency of CRC, the inclusion of CRP measurements from adenoma cases in the Hawaii CCS may have augmented our study's capture of a relevant etiological window for the association between inflammation and CRC.

Another limitation to our study is the use of functional information from GTEx[34] transcriptome data , which was based on tissue samples from the transverse colon but not restricted to the colorectal mucosa, or intestinal stem cells that are etiologically more relevant to risk of CRC[62]. Additionally, CRP concentrations may not correlate well to colonic mucosal inflammation, and in previous studies they have been shown to be falsely low despite active mucosal inflammation[63]. CRP is known to have two isoforms: pentameric CRP (pCRP) and monomeric isoform of CRP (mCRP). The isoform produced by the liver, pCRP, has primarily anti-inflammatory effects and breaks down into the highly pro-inflammatory isoform, mCRP, in tissues with injury or chronic inflammation. Pro-inflammatory effects of mCRP include the

stimulation of inflammatory cytokines and release of reactive oxygen species and these may be more relevant to CRC development. However, the isoform mCRP is not soluble in plasma and is primarily constrained to inflamed tissue. As a result, CRP assays may not necessarily measure the inflammation most relevant for CRC development[64]. A poor correlation between systemic inflammation measured by CRP and mucosal inflammation may have contributed to the inconsistent results. It is also possible that systemic inflammation (as measured by circulating CRP) influences risk of CRC through pathways that are not independent of the risk factors we adjusted for in our study, such as BMI, smoking status, exercise, and postmenopausal hormone use in women. For example, BMI was the strongest predictor for CRP concentration of all adjustment variables examined in our analysis and if the influence of CRP on CRC risk is not independent of BMI then adjustment for BMI would have also controlled for any direct or indirect effects associated with CRP.

CRP is the biomarker most often used to measure systemic inflammation but other biomarkers, or multiple biomarkers rather than a single biomarker, may better evaluate the relationship between inflammation and CRC. Interleukin-6 (IL-6) and tumor necrosis factor (TNF)-α, are also commonly used to measure chronic inflammation. There have been positive associations reported between IL-6 concentrations and colon cancer risk[19,65]. However, since CRP production is stimulated by IL-6, one might also expect a similar ambiguity to CRP results from studies of IL-6. A recent meta-analysis of studies of the inflammatory markers CRP, IL-6 and TNF-α found no associations with colorectal adenoma risk[66]. A positive association between log-transformed IL-6 and CRC risk which varied according to BMI was reported from the Health Professionals Follow-up Study of inflammatory markers[67]. Positive associations reported between adipokines and CRC risk in WHI[68] and EPIC[69] suggest that measures of adipokines and other inflammatory

molecules secreted by adipose tissue into the bloodstream may be more specific than CRP to CRC risk

Aside from these limitations our study had several strengths. We had the largest study sample of this type of analysis to date, with well harmonized individual level data for demographics and CRP measurements, as well as genome-wide genotype data enhanced by imputation. Our novel set-based testing approach, MiSTi, may have increased power to detect interaction effects because it decreased the multiple test burden and incorporated *a priori* functional information[44]. Circulating CRP concentrations provided an objective measure for an observable intermediate phenotype for systemic inflammation. To our knowledge, our study was the first to analyze gene-CRP interaction effects using a set-based approach with variants predictive of tissue-specific gene expression in the colon.

In conclusion, CRP is often used as a measure of systemic inflammation, and a role for inflammation in CRC risk is supported by many previous studies. Our findings may reflect that mechanisms by which systemic inflammation (as measured by CRP) influences CRC risk are not independent from other risk factors for CRC. With a large GWAS consortium built on well-characterized studies, we were well-positioned to identify potential interactions between genes and CRP with respect to CRC risk. Our results indicated that CRP concentrations do not interact with predicted gene expression in the colon to influence CRC risk. Future studies should explore other inflammatory markers with greater specificity that may illuminate the roles of systemic or localized inflammation in CRC development.

Table 3.1 Characteristics of participants in GECCO with CRP measurements

	Cases (n=2,625)	Controls (N=2,798)
CRP (mg/L)		
Mean (SD)	4.15 (6.07)	3.28 (5.09)
Sex		
Female	2,016 (76.8%)	1,876 (67.0%)
Menopausal status		
Premenopausal	74 (3.7%)	95 (5.1%)
Perimenopausal	48 (2.4%)	37 (2.0%)
Postmenopausal	1,873 (92.9%)	1,589 (84.7%)
Missing	21 (1.0%)	155 (8.3%)
Any post-menopausal hormone use		
Yes	651 (34.8%)	728 (45.8%)
Missing	34 (1.8%)	33 (2.1%)
Age (years)		
Mean (SD)	63.8 (8.39)	63.4 (8.19)
BMI (kg/m^2)		
Mean (SD)	27.5 (5.18)	26.6 (4.62)
Missing	15 (0.6%)	21 (0.8%)
Exercise		
Sedentary	1226 (46.7%)	912 (32.6%)
Non-sedentary	589 (22.4%)	469 (16.8%)
Missing	810 (30.9%)	1417 (50.6%)
Diabetes		
Yes	175 (6.7%)	82 (2.9%)
Missing	54 (2.1%)	508 (18.2%)
Regular aspirin/NSAID use		
Yes	644 (24.5%)	634 (22.7%)
Missing	821 (31.3%)	1193 (42.6%)
Smoking status		
Never smoker	1146 (43.7%)	1268 (45.3%)
Former smoker	1112 (42.4%)	1163 (41.6%)
Smoker	253 (9.6%)	284 (10.2%)
Missing	114 (4.3%)	83 (3.0%)
Alcohol use		
nondrinker	1,128 (43.0%)	919 (32.8%)
1-28g/day	1,185 (45.1%)	1,157 (41.4%)
>28g/day	244 (9.3%)	190 (6.8%)
Missing	68 (2.6%)	532 (19.0%)

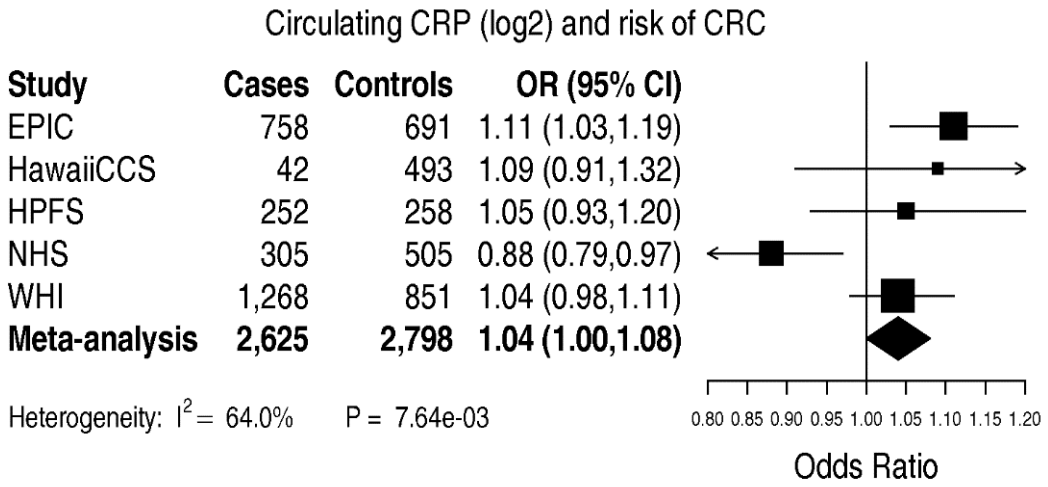
CRP: C-reactive protein; NSAID: Non-steroidal anti-inflammatory; SD: Standard deviation.

Table 3.2 Assay characteristics of C-reactive Protein (CRP) data in GECCO

Study	Design	N		Specimen Source	Method	Intra-assay CV%
EPIC[21]	Cohort	Cases Controls	758 691	Serum	hs-CRP assay (Beckman-Coulter)	2.2%
Hawaii CCS[54]	Case-Control	Cases Controls	42 493	Serum	hs-CRP turbidity assay (Pointe Scientific, Inc.)	3.6%
HPFS[67]	Cohort	Cases Controls	252 258	Serum	hs-CRP immunoturbidimetric assay (Denka Seiken Co.)	7.8%
NHS[52]	Cohort	Cases Controls	305 505	Serum	hs-CRP immunoturbidimetric assay (Denka Seiken Co.)	2.2%
WHI[20,70-85]	Cohort	Cases Controls	1,268 851	Serum; Plasma	hs-CRP on Behring Nephelometer II (Dade Behring, Inc.) ; IMMULITE System (Siemens); and Immunoturbidimetric on Roche analyzer;	3.7%

Abbreviations - CV: Coefficient of variation; EPIC: European Prospective Investigation into Cancer and Nutrition; Hawaii CCS: Hawaiian Adenoma Case Control Study; HPFS: Health Professionals Follow-up Study; NHS: Nurses' Health Study; WHI: Women's Health Initiative.

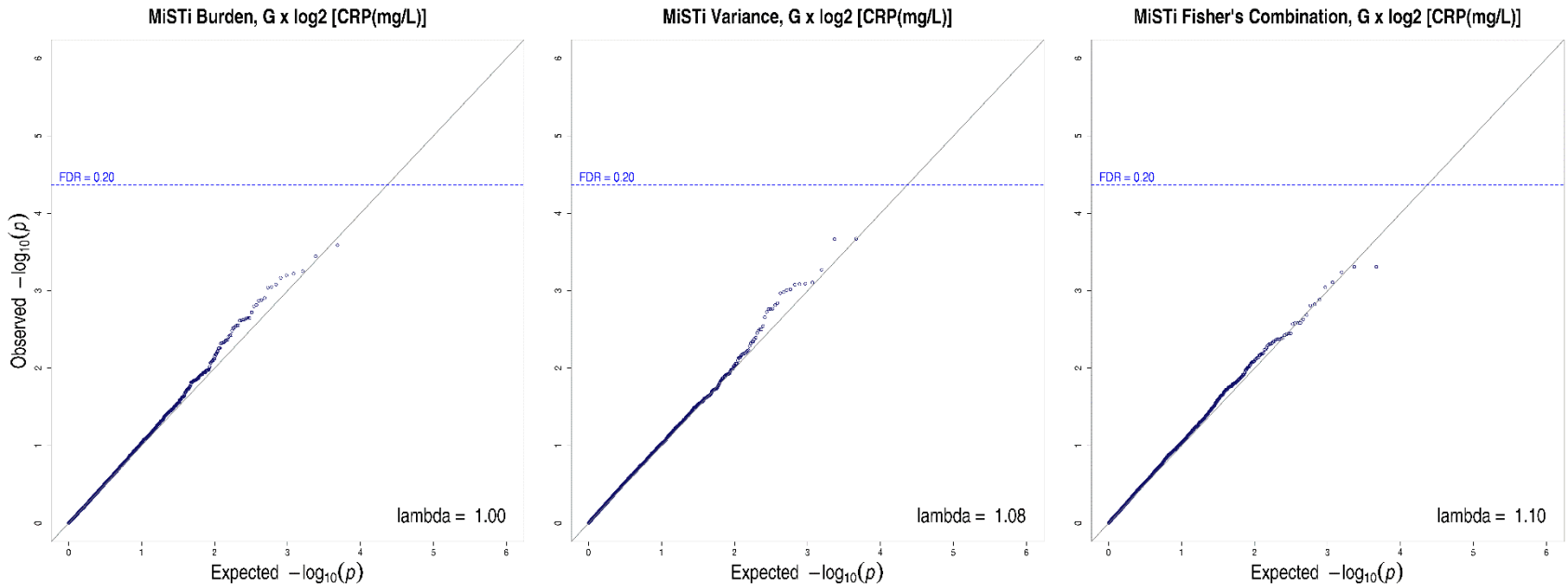
Figure 3.1 Association between circulating C-Reactive Protein (CRP) and risk of CRC



Abbreviations: EPIC: European Prospective Investigation into Cancer and Nutrition; Hawaii CCS: Hawaiian Adenoma Case Control Study; HPFS: Health Professionals Follow-up Study; NHS: Nurses' Health Study; WHI: Women's Health Initiative.

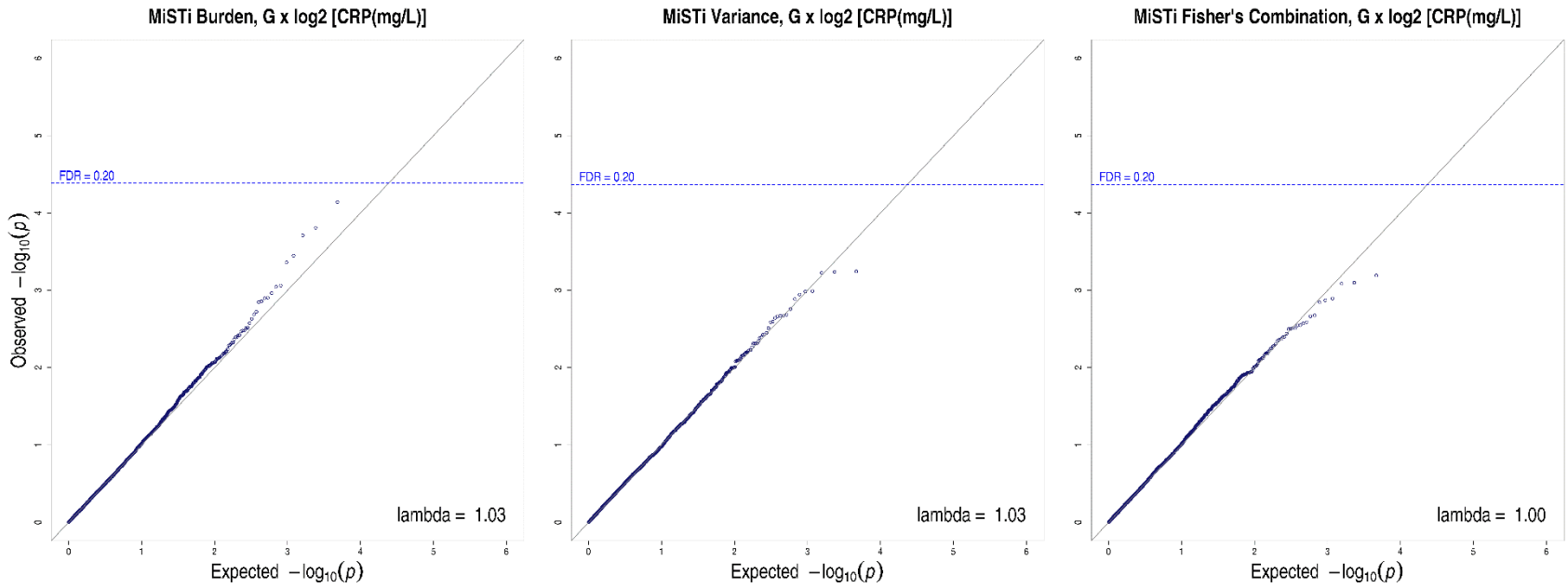
Odds Ratios (ORs) for the association between circulating CRP concentration and risk of colorectal cancer. ORs represent risk per a two-fold higher increase in CRP concentration on the original scale (which corresponds to a difference in log-transformed CRP concentrations of log 2). ORs are adjusted for age at reference, sex, BMI (kg/m2), and post-menopausal hormone replacement use (yes/no). The summary OR is calculated from a fixed-effects meta-analysis with inverse-variance weighting.

Figure 3.2 Q-Q plots from gene-set interaction tests (N=4839) with circulating CRP for risk of CRC



Plots are from results from the MiSTi burden component, MiSTi variance component, and combined MiSTi burden and variance components using Fisher's Combination method. MiSTi models for the interaction between genes and $\log_2$-transformed CRP concentrations were adjusted for age at the reference time, sex, study, genotyping phase, BMI (kg/m^2), post-menopausal hormone replacement use (yes/no), and the principal components to account for potential population substructure. A False Discovery Rate (FDR) threshold of 0.20 is labeled in blue.

Figure 3.3 Q-Q plots from gene-set interaction tests (N=4839) with circulating CRP for risk of CRC excluding the Nurses' Health Study.



Results are from the MiSTi burden component, MiSTi variance component, and combined MiSTi burden and variance components using Fisher's Combination method. MiSTi models for the interaction between genes and $\log_2$-transformed CRP concentrations were adjusted for age at the reference time, sex, study, genotyping phase, BMI (kg/m^2), post-menopausal hormone replacement use (yes/no), and the principal components to account for potential population substructure. A False Discovery Rate (FDR) threshold of 0.20 is labeled in blue.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394-424.

2. Peters U, Bien S, Zubair N. Genetic architecture of colorectal cancer. *Gut.* 2015;64(10):1623-1636.

3. Coussens LM, Werb Z. Inflammation and cancer. *Nature.* 2002;420(6917):860-867.

4. Terzic J, Grivennikov S, Karin E, Karin M. Inflammation and colon cancer. *Gastroenterology.* 2010;138(6):2101-2114 e2105.

5. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. *Cell.* 2010;140(6):883-899.

6. Itzkowitz SH, Yio X. Inflammation and cancer IV. Colorectal cancer in inflammatory bowel disease: the role of inflammation. *Am J Physiol Gastrointest Liver Physiol.* 2004;287(1):G7-17.

7. Chan AT, Giovannucci EL, Meyerhardt JA, Schernhammer ES, Curhan GC, Fuchs CS. Long-term use of aspirin and nonsteroidal anti-inflammatory drugs and risk of colorectal cancer. *JAMA.* 2005;294(8):914-923.

8. Ruder EH, Laiyemo AO, Graubard BI, Hollenbeck AR, Schatzkin A, Cross AJ. Non-steroidal anti-inflammatory drugs and colorectal cancer risk in a large, prospective cohort. *Am J Gastroenterol.* 2011;106(7):1340-1350.

9. Cole BF, Logan RF, Halabi S, et al. Aspirin for the chemoprevention of colorectal adenomas: meta-analysis of the randomized trials. *J Natl Cancer Inst.* 2009;101(4):256-266.

10. Flossmann E, Rothwell PM, British Doctors Aspirin T, the UKTIAAT. Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies. *Lancet.* 2007;369(9573):1603-1613.

11. Ulrich CM, Bigler J, Potter JD. Non-steroidal anti-inflammatory drugs for cancer prevention: promise, perils and pharmacogenetics. *Nat Rev Cancer.* 2006;6(2):130-140.

12. Kim S, Baron JA, Mott LA, et al. Aspirin may be more effective in preventing colorectal adenomas in patients with higher BMI (United States). *Cancer Causes Control.* 2006;17(10):1299-1304.

13. Eberhart CE, Coffey RJ, Radhika A, Giardiello FM, Ferrenbach S, DuBois RN. Up-regulation of cyclooxygenase 2 gene expression in human colorectal adenomas and adenocarcinomas. *Gastroenterology.* 1994;107(4):1183-1188.

14. Kargman SL, O'Neill GP, Vickers PJ, Evans JF, Mancini JA, Jothy S. Expression of prostaglandin G/H synthase-1 and -2 protein in human colon cancer. *Cancer Res.* 1995;55(12):2556-2559.

15. Sano H, Kawahito Y, Wilder RL, et al. Expression of cyclooxygenase-1 and -2 in human colorectal cancer. *Cancer Res.* 1995;55(17):3785-3789.

16. Lumeng CN, Saltiel AR. Inflammatory links between obesity and metabolic disease. *J Clin Invest.* 2011;121(6):2111-2117.

17. Moghaddam AA, Woodward M, Huxley R. Obesity and risk of colorectal cancer: a meta-analysis of 31 studies with 70,000 events. *Cancer Epidemiol Biomarkers Prev.* 2007;16(12):2533-2547.

18. Volanakis JE. Human C-reactive protein: expression, structure, and function. *Mol Immunol.* 2001;38(2-3):189-197.

19. Zhou B, Shu B, Yang J, Liu J, Xi T, Xing Y. C-reactive protein, interleukin-6 and the risk of colorectal cancer: a meta-analysis. *Cancer Causes Control.* 2014;25(10):1397-1405.

20. Toriola AT, Cheng TY, Neuhouser ML, et al. Biomarkers of inflammation are associated with colorectal cancer risk in women but are not suitable as early detection markers. *Int J Cancer.* 2013;132(11):2648-2658.

21. Aleksandrova K, Jenab M, Boeing H, et al. Circulating C-reactive protein concentrations and risks of colon and rectal cancer: a nested case-control study within the European Prospective Investigation into Cancer and Nutrition. *Am J Epidemiol.* 2010;172(4):407-418.

22. Zeng C, Matsuda K, Jia WH, et al. Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. *Gastroenterology.* 2016;150(7):1633-1645.

23. Wang H, Burnett T, Kono S, et al. Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat Commun.* 2014;5:4613.

24. Schumacher FR, Schmit SL, Jiao S, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun.* 2015;6:7138.

25. Schmit SL, Schumacher FR, Edlund CK, et al. A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. *Carcinogenesis.* 2014;35(11):2512-2519.

26. Peters U, Jiao S, Schumacher FR, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology.* 2013;144(4):799-807 e724.

27. Al-Tassan NA, Whiffin N, Hosking FJ, et al. A new GWAS and meta-analysis with

1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci Rep.* 2015;5:10442.

28. Real LM, Ruiz A, Gayan J, et al. A colorectal cancer susceptibility new variant at 4q26 in the Spanish population identified by genome-wide association analysis. *PLoS One.* 2014;9(6):e101178.

29. 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.

30. Schmit SL, Edlund CK, Schumacher FR, et al. Novel Common Genetic Susceptibility Loci for Colorectal Cancer. *J Natl Cancer Inst.* 2019;111(2):146-157.

31. Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer.* 2002;99(2):260-266.

32. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med.* 2000;343(2):78-85.

33. Thomas D. Gene--environment-wide association studies: emerging approaches. *Nat Rev Genet.* 2010;11(4):259-272.

34. Aguet F, Brown AA, Castel SE, et al. Genetic effects on gene expression across human tissues. *Nature.* 2017;550(7675):204-213.

35. Curran-Everett D. Explorations in statistics: the log transformation. *Adv Physiol Educ.* 2018;42(2):343-347.

36. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. *Am J Hum Genet.* 2013;93(4):687-696.

37. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48(10):1284-1287.

38. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* 2016;48(10):1279-1283.

39. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.

40. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet.* 2015;47(9):1091-1098.

41. The Genomes Project C, Auton A, Abecasis GR, et al. A global reference for human genetic variation. *Nature.* 2015;526:68.

42. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations inR. *Journal of Statistical Software.* 2011;45(3).

43. Team RC. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing;2013.

44. Su YR, Di CZ, Hsu L, Genetics, Epidemiology of Colorectal Cancer C. A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics.* 2017;18(1):119-131.

45. Jiao S, Peters U, Berndt S, et al. Estimating the heritability of colorectal cancer. *Hum Mol Genet.* 2014;23(14):3898-3905.

46. Fisher RA GS, Britain G, Généticien S. *Statistical methods for research workers.* Vol 14. Edinburgh: Oliver and Boyd; 1970.

47. Trichopoulos D, Psaltopoulou T, Orfanos P, Trichopoulou A, Boffetta P. Plasma C-reactive protein and risk of cancer: a prospective study from Greece. *Cancer Epidemiol Biomarkers Prev.* 2006;15(2):381-384.

48. Siemes C, Visser LE, Coebergh JW, et al. C-reactive protein levels, variation in the C-reactive protein gene, and cancer risk: the Rotterdam Study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* 2006;24(33):5216-5222.

49. Otani T, Iwasaki M, Sasazuki S, Inoue M, Tsugane S, Japan Public Health Center-Based Prospective Study G. Plasma C-reactive protein and risk of colorectal cancer in a nested case-control study: Japan Public Health Center-based prospective study. *Cancer Epidemiol Biomarkers Prev.* 2006;15(4):690-695.

50. Ito Y, Suzuki K, Tamakoshi K, et al. Colorectal cancer and serum C-reactive protein levels: a case-control study nested in the JACC Study. *Journal of epidemiology / Japan Epidemiological Association.* 2005;15 Suppl 2:S185-189.

51. Zhang SM, Buring JE, Lee IM, Cook NR, Ridker PM. C-reactive protein levels are not associated with increased risk for colorectal cancer in women. *Annals of internal medicine.* 2005;142(6):425-432.

52. Chan AT, Ogino S, Giovannucci EL, Fuchs CS. Inflammatory markers are associated with risk of colorectal cancer and chemopreventive response to anti-inflammatory drugs. *Gastroenterology.* 2011;140(3):799-808, quiz e711.

53. Tsilidis KK, Branchini C, Guallar E, Helzlsouer KJ, Erlinger TP, Platz EA. C-reactive protein and colorectal cancer risk: a systematic review of prospective studies. *Int J Cancer.* 2008;123(5):1133-1140.

54. Ognjanovic S, Yamamoto J, Saltzman B, et al. Serum CRP and IL-6, genetic variants and risk of colorectal adenoma in a multiethnic population. *Cancer Causes Control.*

2010;21(7):1131-1138.

55. Gabay C, Kushner I. Acute-phase proteins and other systemic responses to inflammation. *N Engl J Med.* 1999;340(6):448-454.

56. Ockene IS, Matthews CE, Rifai N, Ridker PM, Reed G, Stanek E. Variability and classification accuracy of serial high-sensitivity C-reactive protein measurements in healthy adults. *Clin Chem.* 2001;47(3):444-450.

57. Meier-Ewert HK, Ridker PM, Rifai N, Price N, Dinges DF, Mullington JM. Absence of diurnal variation of C-reactive protein concentrations in healthy human subjects. *Clin Chem.* 2001;47(3):426-430.

58. Mantovani A, Allavena P, Sica A, Balkwill F. Cancer-related inflammation. *Nature.* 2008;454(7203):436-444.

59. Brenner H, Altenhofen L, Katalinic A, Lansdorp-Vogelaar I, Hoffmeister M. Sojourn time of preclinical colorectal cancer by sex and age: estimates from the German national screening colonoscopy database. *Am J Epidemiol.* 2011;174(10):1140-1146.

60. Brenner H, Hoffmeister M, Stegmaier C, Brenner G, Altenhofen L, Haug U. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. *Gut.* 2007;56(11):1585-1589.

61. Rutter CM, Knudsen AB, Marsh TL, et al. Validation of Models Used to Inform Colorectal Cancer Screening Guidelines: Accuracy and Implications. *Med Decis Making.* 2016;36(5):604-614.

62. Gehart H, Clevers H. Tales from the crypt: new insights into intestinal stem cells. *Nat Rev Gastroenterol Hepatol.* 2019;16(1):19-34.

63. Chang S, Malter L, Hudesman D. Disease monitoring in inflammatory bowel disease. *World J Gastroenterol.* 2015;21(40):11246-11259.

64. Del Giudice M, Gangestad SW. Rethinking IL-6 and CRP: Why they are more than inflammatory biomarkers, and why it matters. *Brain Behav Immun.* 2018;70:61-75.

65. Kakourou A, Koutsioumpa C, Lopez DS, et al. Interleukin-6 and risk of colorectal cancer: results from the CLUE II cohort and a meta-analysis of prospective studies. *Cancer Causes Control.* 2015;26(10):1449-1460.

66. Zhang X, Liu S, Zhou Y. Circulating levels of C-reactive protein, interleukin-6 and tumor necrosis factor-alpha and risk of colorectal adenomas: a meta-analysis. *Oncotarget.* 2016;7(39):64371-64379.

67. Song M, Wu K, Ogino S, Fuchs CS, Giovannucci EL, Chan AT. A prospective study of plasma inflammatory markers and risk of colorectal cancer in men. *Br J Cancer.* 2013;108(9):1891-1898.

68. Ho GY, Wang T, Gunter MJ, et al. Adipokines linking obesity with colorectal cancer risk in postmenopausal women. *Cancer Res.* 2012;72(12):3029-3037.

69. Aleksandrova K, Jenab M, Bueno-de-Mesquita HB, et al. Biomarker patterns of inflammatory and metabolic pathways are associated with risk of colorectal cancer: results from the European Prospective Investigation into Cancer and Nutrition (EPIC). *Eur J Epidemiol.* 2014;29(4):261-275.

70. Wassertheil-Smoller S, Hendrix SL, Limacher M, et al. Effect of estrogen plus progestin on stroke in postmenopausal women: the Women's Health Initiative: a randomized trial. *JAMA.* 2003;289(20):2673-2684.

71. Rajpathak SN, Kaplan RC, Wassertheil-Smoller S, et al. Resistin, but not adiponectin and leptin, is associated with the risk of ischemic stroke among postmenopausal women: results from the Women's Health Initiative. *Stroke.* 2011;42(7):1813-1820.

72. Pradhan AD, Manson JE, Rossouw JE, et al. Inflammatory biomarkers, hormone replacement therapy, and incident coronary heart disease: prospective analysis from the Women's Health Initiative observational study. *JAMA.* 2002;288(8):980-987.

73. Liu S, Tinker L, Song Y, et al. A prospective study of inflammatory cytokines and diabetes mellitus in a multiethnic cohort of postmenopausal women. *Arch Intern Med.* 2007;167(15):1676-1685.

74. Wang L, Manson JE, Gaziano JM, et al. Plasma adiponectin and the risk of hypertension in white and black postmenopausal women. *Clin Chem.* 2012;58(10):1438-1445.

75. Birmann BM, Neuhouser ML, Rosner B, et al. Prediagnosis biomarkers of insulin-like growth factor-1, insulin, and interleukin-6 dysregulation and multiple myeloma risk in the Multiple Myeloma Cohort Consortium. *Blood.* 2012;120(25):4929-4937.

76. Lee IM, Sesso HD, Ridker PM, Mouton CP, Stefanick ML, Manson JE. Physical activity and inflammation in a multiethnic cohort of women. *Med Sci Sports Exerc.* 2012;44(6):1088-1096.

77. Cook NR, Paynter NP, Eaton CB, et al. Comparison of the Framingham and Reynolds Risk scores for global cardiovascular risk prediction in the multiethnic Women's Health Initiative. *Circulation.* 2012;125(14):1748-1756, S1741-1711.

78. Kaplan RC, McGinn AP, Baird AE, et al. Inflammation and hemostasis biomarkers for predicting stroke in postmenopausal women: the Women's Health Initiative Observational Study. *J Stroke Cerebrovasc Dis.* 2008;17(6):344-355.

79. Gray SL, LaCroix AZ, Aragaki AK, et al. Angiotensin-converting enzyme inhibitor use and incident frailty in women aged 65 and older: prospective findings from the Women's Health Initiative Observational Study. *J Am Geriatr Soc.* 2009;57(2):297-303.

80. Mares JA, LaRowe TL, Snodderly DM, et al. Predictors of optical density of lutein and zeaxanthin in retinas of older women in the Carotenoids in Age-Related Eye Disease Study, an ancillary study of the Women's Health Initiative. *Am J Clin Nutr.* 2006;84(5):1107-1122.

81. Vilasdechanon N, Kaewchur T, Ua-Apisitwong S. Radioiodine dosimetry in advance differentiated thyroid carcinoma with poor clinical outcome: A preliminary report. *J Nucl Med.* 2012;53.

82. Horn LV, Tian L, Neuhouser ML, et al. Dietary patterns are associated with disease risk among participants in the Women's Health Initiative Observational Study. *J Nutr.* 2012;142(2):284-291.

83. Walitt B, Mackey R, Kuller L, et al. Predictive value of autoantibody testing for validating self-reported diagnoses of rheumatoid arthritis in the Women's Health Initiative. *Am J Epidemiol.* 2013;177(9):887-893.

84. Rohan TE, Heo M, Choi L, et al. Body fat and breast cancer risk in postmenopausal women: a longitudinal study. *J Cancer Epidemiol.* 2013;2013:754815.

85. Coviello AD, Haring R, Wellons M, et al. A genome-wide association meta-analysis of circulating sex hormone-binding globulin reveals multiple Loci implicated in sex steroid hormone regulation. *PLoS Genet.* 2012;8(7):e1002805.

CHAPTER 4. CONCLUSION

In this project, we aimed to better understand the relationships and underlying mechanisms of dietary factors, chronic inflammation to CRC risk due to heritable genetic variation. For this endeavor, we conducted extensive GxE analyses using predicted gene expression in the normal colon tissue to weight and focus our GxE testing. To our knowledge, these were the first and largest studies to analyze the interactions between dietary factors, and the CRP biomarker, with predicted tissue-specific gene expression and CRC risk. We found a novel gene interaction with processed meat with a plausible biological mechanism for interaction with gene expression in the colon. Our findings suggest that previously hypothesized mechanisms by which red meats and processed meats influence colorectal cancer such as oxidative stress and inflammation could indeed underlie the association between increased intake of processed meat and increased risk of colorectal cancer. Furthermore, our findings suggest that dietary intake can indeed interact with genotype. Our findings also indicated that CRP concentrations do not interact with predicted gene expression in the colon to influence CRC risk. Future studies should explore other inflammatory markers with greater specificity that may illuminate the roles of systemic or localized inflammation in CRC development.

APPENDIX A: SUPPLEMENTAL TEXTS

Supplemental Text 1. Description of Studies for Chapter 2

Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC)

The ATBC Study was conducted in Finland as a joint project between the National Institute for Health and Welfare of Finland and the US National Cancer Institute (NCI). The overall design, rationale, objectives, and initial results of this intervention study have been published[1,2]. Briefly, this was a randomized, double-blind, placebo-controlled, primary prevention trial to determine whether daily supplementation with alpha-tocopherol, beta-carotene, or both would reduce the incidence of lung or other cancers among male smokers. A total of 29,133 men between the ages of 50 and 69 years, who smoked at least five cigarettes per day, were recruited from southwestern Finland between 1985 and 1988, and randomly assigned to one of four groups based on a 2×2 factorial design. Men who had prior cancer or serious illness or who reported current use of vitamins E (>20mg/day), A (>20,000 IU/day), or beta-carotene (>6 mg/day) were ineligible. Participants received either alpha-tocopherol (50 mg/day) as dl-alpha-tocopheryl acetate, beta-carotene (20 mg/day) as all-trans-beta-carotene, both supplements, or placebo capsules for 5-8 years (median 6.1 years) until death or trial closure (April 30, 1993). Data Collection: At baseline, study subjects completed a general risk factor, smoking, and medical history questionnaire, along with a food frequency (use) questionnaire, which consisted of a modified diet history, including both portion size and frequency of consumption for 203 food items and 73 mixed dishes. This instrument was intended to measure usual consumption over the previous 12 months. Nutrient intake was estimated using food composition data available from the National Institute for Health and Welfare of Finland. Height, weight, blood pressure, heart rate, and visual acuity were measured. Follow-up consisted of three visits annually to the local field center, during which the men were asked about their health, use of non-trial vitamin supplements, and smoking habits since the last visit. Height, weight, blood pressure, heart rate,

and visual acuity were measured once a year. At 3 years, the food frequency questionnaire was repeated for all participants. Participants in the study were asked to contact their local study center as soon as possible if they were diagnosed with cancer, and they were then invited for a follow-up visit, where they completed another food frequency questionnaire.

American Cancer Society Cancer Prevention Study II (CPS-II)

As described previously, CPS-II is a cohort study started by the American Cancer Society in 1982 to investigate the relationship between dietary, lifestyle and other etiologic factors and cancer mortality[3]. Approximately 1.2 million men and women were enrolled in the study from 50 states in the U.S. In 1992, a subset of these participants (N~184,000) were enrolled in the CPS-II Nutrition Cohort to examine the relationship between dietary and other exposures and cancer incidence. Blood samples were drawn from approximately 39,376 members of the Nutrition Cohort from 1998 to 2001, and buccal cells were collected from 69,467 additional members from 2001 to 2002. Cancer cases are identified by self-report through biennial follow-up questionnaires or through linkage with the National Death Index, followed by verification through medical records or linkage to state cancer registries. A total of 548 men and women diagnosed with colon or rectal cancer after providing a blood or buccal cell sample were genotyped for this study. Population-based control participants genotyped for this study included 538 men and women from the CPS-II Nutrition Cohort, individually matched to a case on sex, race/ethnicity, date of birth, date of sample collection, and DNA source (blood or buccal cell).

Association STudy Evaluating RISK for sporadic colorectal cancer (ASTERISK)

Participants were recruited from the Pays de la Loire region in France between December 2002 and March 2006[4] . Eligibility criteria for cases included being of Caucasian origin, being greater than or 40 years of age at diagnosis and having no family history of colorectal cancer or polyps. Cases were patients with first primary colorectal cancer diagnosed in one of the six public

hospitals and five clinics located in the Pays de la Loire region which participated in the study. Cases were confirmed based on medical and pathology reports. Controls were recruited at two Health Examination Centers of the Pays de la Loire region, and the recruitment of controls greater than or 70 years was completed in the departments of internal medicine and hepatogastroenterology of the University Hospital Center of Nantes, located in the same region. Controls were eligible to participate if they were Caucasian, aged greater than or 40 years, and had no family history of colorectal cancer or polyps. In the presence of the physician, each participant filled out a standardized questionnaire on family information, medical history, lifestyle, and dietary intake. Cases and controls provided a blood sample.

Colon Cancer Family Registry (CCFR)

The CCFR is an NCI-supported consortium consisting of six centers dedicated to the establishment of a comprehensive collaborative infrastructure for interdisciplinary studies in the genetic epidemiology of colorectal cancer[5]. The CCFR includes data from approximately 30,500 total subjects (10,500 probands, and 20,000 unaffected and affected relatives and unrelated controls). Cases and controls, age 20 to 74 years, were recruited at the six participating centers beginning in 1998. CCFR implemented a standardized questionnaire that is administered to all participants, and includes established and suspected risk factors for colorectal cancer, which includes questions on medical history and medication use, reproductive history (for female participants), family history, physical activity, demographics, alcohol and tobacco use, and dietary factors. The Set 1 scan, which has been described previously[6], includes population-based cases and age-matched controls from the three population-based centers: Seattle, Toronto and Australia. Cases were genetically enriched by over-sampling those with an early age at onset or positive family history. Controls were matched to cases on age and sex. All cases and controls were self-reported as White, which was confirmed with genotype data. The Set 2 scan includes

population-based cases and matched controls from all six Colon CFR centers including Mayo Clinic, Hawaii Cancer Registry, University of Southern California, Fred Hutchinson Cancer Research Center, Cancer Care Ontario and University of Melbourne. As with Set 1, cases were genetically enriched by over-sampling those with an early age at onset or positive family history. Controls were same generation family controls.

Colorectal Cancer Genetics & Genomics, Spanish study (CRCGEN)

CRCGEN combines data of three case-control studies. The first one, performed in University Hospital of Bellvitge, L'Hospitalet, Barcelona, recruited 304 incident and pathology confirmed, CRC cases and 293 age and sex frequency-matched hospital controls during the period 1996-1998. The control group consisted of patients without previous colorectal cancer who had been randomly selected among those admitted to the same hospital during the same period. To avoid selection bias, the criterion of inclusion in the control group was a new diagnosis. The second study, performed in the same hospital during the period 2007-2015, included a total of 324 cases and 376 population controls. The control group was recruited by inviting to participate subjects selected from the primary health care lists of the hospital's referral area, frequency matched by age and sex. The third study was conducted in Hospital of Leon, Leon, during 2008-2013. A total of 325 incident CRC cases and 407 population controls were included. The control population was recruited by inviting to participate subjects selected from the primary health care lists, frequency matched by age and sex. Written informed consent was required from all participants. Each Hospital's ethics committees (Bellvitge and Leon) approved the protocols of the study.

Darmkrebs: Chancen der Verhütung durch Screening (DACHS)

This German study was initiated as a large population-based case-control study in 2003 in the Rhine-Neckar-Odenwald region (southwest region of Germany) to assess the potential of endoscopic screening for reduction of colorectal cancer risk and to investigate etiologic

determinants of disease, particularly lifestyle/environmental factors and genetic factors[7,8]. Cases with a first diagnosis of invasive colorectal cancer (International Classification of Diseases 10 codes C18-C20) who were at least 30 years of age (no upper age limit), German speaking, a resident in the study region, and mentally and physically able to participate in a one-hour interview, were recruited by their treating physicians either in the hospital a few days after surgery, or by mail after discharge from the hospital. Cases were confirmed based on histologic reports and hospital discharge letters following diagnosis of colorectal cancer. All hospitals treating colorectal cancer patients in the study region participated. Based on estimates from population-based cancer registries, more than 50% of all potentially eligible patients with incident colorectal cancer in the study region were included. Community-based controls were randomly selected from population registries, employing frequency matching with respect to age (5-year groups), sex, and county of residence. Controls with a history of colorectal cancer were excluded. Controls were contacted by mail and follow-up calls. The participation rate was 51%. During an in-person interview, data were collected on demographics, medical history, family history of CRC, and various life-style factors, as were blood and mouthwash samples. This analysis includes participants recruited up to 2010 in this ongoing study, controls were frequency matched to cases on age, gender, and county of residence.

Diet, Activity, and Lifestyle Study (DALS)

DALS is a population-based case-control study of colon cancer[9]**.**. Participants were recruited between 1991 and 1994 from three locations: the Kaiser Permanente Medical Care Program (KPMCP) of Northern California, an eight-county area in Utah, and the metropolitan Twin Cities area of Minnesota. Eligibility criteria for cases included age at diagnosis between 30 and 79 years, diagnosis with first primary colon cancer (International Classification of Diseases for Oncology-2 codes 18.0 and 18.2-18.9) between October 1, 1991 and September 30, 1994,

English speaking, and competency to complete the interview. Individuals with cancer of the rectosigmoid junction or rectum were excluded, as were those with a pathology report noting familial adenomatous polyposis, Crohn's disease, or ulcerative colitis. A rapid-reporting system was used to identify all incident cases of colon cancer resulting in the majority of cases being interviewed within four months of diagnosis. Controls from KPMCP were randomly selected from membership lists. In Utah, controls under 65 years of age were randomly selected through random-digit dialing and driver license lists. Controls, 65 years of age and older, were randomly selected from Health Care Financing Administration lists. In Minnesota, controls were identified from Minnesota driver's license or state ID lists. Controls were matched to cases by 5-year age groups and sex. The Set I scan consisted of a subset of the study designed above, from Utah, Minnesota, and KPMCP, and was restricted to subjects who self-reported as White non-Hispanic. The Set 2 scan consisted of subjects from Utah and Minnesota that were not genotyped in Set 1. Set 2 was restricted to subjects who self-reported as White non-Hispanic and those that had appropriate consent to post data to dbGaP.

Hawai'i Colorectal Cancer Studies 2 & 3 (Colo2&3)

Patients with colorectal cancer were identified through the rapid reporting system of the Hawaii SEER registry and consisted of all Japanese, Caucasian, and Native Hawaiian residents of Oahu who were newly diagnosed with an adenocarcinoma of the colon or rectum between January 1994 and August 1998[10]. Control subjects were selected from participants in an on-going population-based health survey conducted by the Hawaii State Department of Health and from Health Care Financing Administration participants. Controls were matched to cases by sex, ethnicity, and age (within two years). Personal interviews were obtained from 768 matched pairs, resulting in a participation rate of 58.2% for cases and 53.2% for controls. A questionnaire, administered during an in-person interview, included questions about demographics, lifetime

history of tobacco, alcohol use, aspirin use, physical activity, personal medical history, family history of colorectal cancer, height and weight, diet (Food Frequency Questionnaire), and postmenopausal hormone use. A blood sample was obtained from 548 (71%) of interviewed cases and 662 (86%) of interviewed controls. SEER staging information was extracted from the Hawaii Tumor Registry. In GECCO, self-reported Caucasian subjects with DNA, and clinical and epidemiologic data were selected for genotyping.

Health Professionals Follow-up Study (HPFS)

The HPFS is a parallel prospective study to the Nurses' Health Study (NHS)[11]. The HPFS cohort comprises 51,529 men who, in 1986, responded to a mailed questionnaire. The participants are U.S. male dentists, optometrists, osteopaths, podiatrists, pharmacists, and veterinarians born between 1910 and 1946. Participants have provided information on health-related exposures, including: current and past smoking history, age, weight, height, diet, physical activity, aspirin use, and family history of colorectal cancer. Colorectal cancer and other outcomes were reported by participants or next-of-kin and followed up through review of the medical and pathology record by physicians. Overall, more than 97% of self-reported colorectal cancers were confirmed by medical record review. Information was abstracted on histology and primary location. Incident cases are defined as those occurring after the subject provided the blood sample. Prevalent cases are defined as those occurring after enrollment in the study, but prior to the subject providing the blood sample. Follow-up has been excellent, with 94% of the men responding to date. Colorectal cancer cases were ascertained through January 1, 2008. In 1993-95, 18,825 men in HPFS mailed in blood samples by overnight courier which were aliquoted into buffy coat and stored in liquid nitrogen. In 2001-04, 13,956 men in HPFS who had not previously provided a blood sample mailed in a "swish-and-spit" sample of buccal cells. Incident cases are defined as those occurring after the subject provided a blood or buccal sample.

Prevalent cases are defined as those occurring after enrollment in the study in 1986, but prior to the subject providing either a blood or buccal sample. After excluding participants with histories of cancer (except non-melanoma skin), ulcerative colitis, or familial polyposis, two case-control sets were constructed from which DNA was isolated from either buffy coat or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the cases; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the case. For both case-control sets, matching criteria included year of birth (within 1 year) and month/year of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s).

In addition to colorectal cancer cases and controls, a set of adenoma cases and matched controls with available DNA from buffy coat were selected for genotyping. Over follow-up, data were collected on endoscopic screening practices and, if individuals have been diagnosed with polyp, the polyps were confirmed to be adenomatous by medical record review. Adenoma cases were ascertained through January 1, 2008. A separate case-control set was constructed of participants diagnosed with advanced adenoma matched to control participants who underwent a lower endoscopy in the same period and did not have an adenoma. Advanced adenoma was defined as an adenoma >=1 cm in diameter and / or with tubulovillous, villous, or high-grade dysplasia / carcinoma-in-situ histology. Matching criteria included year of birth (within one year) and month/year of blood sampling (within six months), the reason for their lower endoscopy (screening, family history, or symptoms) and the period of any prior endoscopy (within two

years). Controls matched to cases with a distal adenoma either had a negative sigmoidoscopy or colonoscopy exam and controls matched to cases with proximal adenoma all had a negative colonoscopy.

Kentucky Case-Control Study

The Kentucky Case-Control study was initiated in July 2003 through the University of Kentucky Cancer Center[12,13]. A web-based reporting system implemented by the Kentucky Cancer Registry in 2003 has facilitated rapid report of cases statewide, with approximately 76.8% of all cases reported to the registry within 6 months of diagnosis. Cases (>21 years) diagnosed with histologically confirmed colon cancer and entered into the registry within 6 months of their diagnoses are invited to join the study. Population-based unrelated controls are recruited through random digit dialing and are frequency matched to the cases by age (±5 years), gender, and race. Excluded from the study are those individuals who have been diagnosed with colon cancer because of known hereditary forms of colon cancer or polyposis such as familial adenomatous polyposis (FAP), hereditary non-polyposis colorectal cancer (HNPCC), Peutz-Jeghers, and Cowden disease. Currently there are more than 1,040 incident population-based cases of colorectal cancer and 1,750 population-based controls fully recruited, with comprehensive epidemiologic data, pathology data, and DNA from cases and controls.

Melbourne Collaborative Cohort Study (MCCS)

The MCCS includes both men and women volunteers, aged 40-69 and recruited from the Melbourne metropolitan area in the early 1990s[14]. To recruit a sample with an increased range of dietary exposures, it was decided to deliberately enrich the cohort with migrants to Melbourne from Italy and Greece. The baseline questionnaire included questions on personal medical history and family history of common diseases. Other important environmental variables were also accounted for. Blood samples were collected from all subjects in 15ml lithium heparin

vacutainers. Total plasma cholesterol and glucose were measured immediately using Kodak Ektachem DT60 desktop analysers. A total of 16,962 men and 24,286 women aged between 40 and 69 years were recruited into the cohort between 1990 and 1994[15].

Molecular Epidemiology of Colorectal Cancer (MECC) Study

The Molecular Epidemiology of Colorectal Cancer Study (MECC) is a population-based case-control study of colorectal cancer (CRC)[16]. Incident, pathologically-confirmed CRC cases and controls were recruited from a specific region of northern Israel. Newly-diagnosed CRC cases beginning March 31, 1998, who agreed to participate were interviewed, gave a venous blood sample, and provided permission for tumor tissue retrieval. Written, informed consent was obtained according to Institutional Review Board-approved protocols at Carmel Medical Center in Haifa and the University of Southern California (HS-12-00324, HS-12-00672, and HS-08-00378). Germline DNA was extracted from whole blood for genotyping. The analytic dataset from the MECC study genotyped on the OncoArray and included in the CORECT Phase 2 European GWAS consisted of 3,591 cases of pathologically-confirmed adenocarcinoma and 2,848 controls. In addition, previously genotyped cases and controls were included in the Phase 1 GWAS: these consisted of 484 cases and 498 controls genotyped on the Illumina Omni 2.5 array, and 1,120 cases and 820 controls were genotyped on the Affymetrix Axiom CORECT Set array. Thus, the total number of cases and controls from the MECC study included in Phases 1 and 2 (after quality control for genotyping) was 5,195 cases and 4,166 controls.

Multiethnic Cohort Study (MEC)

MEC was initiated in 1993 to investigate the impact of dietary and environmental factors on major chronic diseases, particularly cancer, in ethnically diverse populations in Hawai'i and California[17]. The study recruited 96,810 men and 118,441 women aged 45 to 75 years between 1993 and 1996. Incident colorectal cancer cases occurring since January 1995, and controls were

contacted for blood or saliva samples. The median interval between diagnosis and blood draw was 14 months (interquartile range, 10-19) among cases and the participation rate 74%. A sample of cohort participants was randomly selected to serve as controls at the onset of the nested case-control study (participation rate 66%). The selection was stratified by sex, age, and race/ethnicity. Colorectal cancer cases are identified through the Rapid Reporting System of the Hawai'i Tumor Registry and through quarterly linkage to the Los Angeles County Cancer Surveillance Program. Both registries are members of SEER. In GECCO, self-reported White subjects from the nested case-control study described above with DNA, and clinical and epidemiologic data were selected for genotyping.

Newfoundland Familial Colon Cancer Registry (NFCCR)

The NFCCR is a case-control study, which includes pathology confirmed CRC cases, less than 75 years of age, diagnosed between January 1, 1999 and December 31, 2003, identified from the Newfoundland Cancer Registry[18]. The Newfoundland Cancer Registry registers all cases of invasive cancer diagnosed among residents of the province of Newfoundland and Labrador. Consenting patients received a family history questionnaire and were asked to provide a blood sample and to permit access to tumor tissue and medical records. If a patient was deceased, we sought the participation of a close relative for the purposes of obtaining the family history and for permission to access tissue blocks and medical records. Use of proxies in this way removes the bias of excluding advanced-stage cancer patients who die before they can give consent. Controls were identified by random digit dialing from the residents of the province and matched to the cases on sex and five-year age group. Controls provided a blood sample and filled out a risk factor questionnaire.

Nurses' Health Study (NHS)

The NHS cohort began in 1976 when 121,700 married female registered nurses aged 30 to 55 years returned the initial questionnaire that ascertained a variety of important health-related exposures[19]. Since 1976, follow-up questionnaires have been mailed every two years. Colorectal cancer and other outcomes were reported by participants or next-of-kin and followed up through review of the medical and pathology record by physicians. Overall, more than 97% of self-reported colorectal cancers were confirmed by medical-record review. Information was abstracted on histology and primary location. Follow-up has been high: as a proportion of the total possible follow-up time, follow-up has been over 92%. Colorectal cancer cases were ascertained through June 1, 2008. In 1989-90, 32,826 women in NHS I, mailed in blood samples by overnight courier which were aliquoted into buffy coat and stored in liquid nitrogen. In 2001-04, 29,684 women in NHS I who did not previously provide a blood sample mailed in a "swish-and-spit" sample of buccal cells. Incident cases are defined as those occurring after the subject provided a blood or buccal sample. Prevalent cases are defined as those occurring after enrollment in the study in 1976, but prior to the subject providing either a blood or buccal sample. After excluding participants with histories of cancer (except non-melanoma skin), ulcerative colitis, or familial polyposis, two case-control sets were constructed from which DNA was isolated from either buffy coat or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the case; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the cases. For both case-control sets, matching criteria included year of

birth (within one year) and month / year of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s).

In addition to colorectal cancer cases and controls, a set of adenoma cases and matched controls with available DNA from buffy coat were selected for genotyping. Over follow-up, data were collected on endoscopic screening practices and, if individuals have been diagnosed with polyp, the polyps confirmed to be adenomatous by medical record review. Adenoma cases were ascertained through June 1, 2008. A separate case-control set was constructed of participants diagnosed with advanced adenoma matched to control participants who underwent a lower endoscopy in the same period and did not have an adenoma. Advanced adenoma was defined as an adenoma > 1 cm in diameter and / or with tubulovillous, villous, or high-grade dysplasia / carcinoma-in-situ histology. Matching criteria included year of birth (within one year) and month/year of blood sampling (within six months), the reason for their lower endoscopy (screening, family history, or symptoms) and the period of any prior endoscopy (within two years). Controls matched to cases with a distal adenoma either had a negative sigmoidoscopy or colonoscopy exam and controls matched to cases with proximal adenoma all had a negative colonoscopy.

Ontario Familial Colorectal Cancer Registry (OFCCR)

For this GECCO analysis, a subset of the Assessment of Risk in Colorectal Tumours in Canada (ARCTIC) from the Ontario Registry for Studies of Familial Colorectal Cancer (OFCCR) was used[20]. Details on the case-control study[21] and the OFCCR, as well as the GWAS results[22] have previously been reported. In brief, cases were confirmed incident colorectal cancer (CRC) cases aged 20 to 74 years, residents of Ontario identified through comprehensive registry and diagnosed between July 1997 and June 2000. Population-based controls were randomly selected among Ontario residents (random-digit-dialing and listing of all Ontario residents) and matched

by sex and 5-year age groups. A total of 1,236 CRC cases and 1,223 controls were successfully genotyped on at least one of the Illumina 1536 GoldenGate assay (Illumina, Inc, San Diego, CA), the Affymetrix GeneChip® Human Mapping 100K and 500K Array Set (Affymetrix, Inc, Santa Clara, CA), and a 10K non-synonymous variant chip. Analysis was based on a set of unrelated subjects who were non-Hispanic, White by self-report or by investigation of genetic ancestry. Further exclusions were made for sample swaps, missing epidemiologic questionnaire data, appendix tumor, or if a subject overlapped with the Colon Cancer Family Registry. Additionally, only samples genotyped on the Affymetrix GeneChip® 500K Array were utilized to avoid convergence issues in imputation.

Physician's Health Study (PHS)

The PHS was established as a randomized, double-blind, placebo-controlled trial of aspirin and ß-carotene among 22,071 healthy U.S. male physicians, between 40 and 84 years of age in 1982[23,24]. Participants completed two mailed questionnaires before being randomly assigned, additional questionnaires at six and 12 months, and questionnaires annually thereafter. In addition, participants were sent postcards at six months to ascertain status. From August 1982 to December 1984, 14,916 baseline blood samples were collected from the physicians during the run-in phase before randomization. When participants report a diagnosis of cancer, medical records and pathology reports are reviewed by study physicians who are blinded to exposure data. Among those who provided baseline blood samples, colorectal cases were ascertained through March 31, 2008, and controls were matched on age (within one year for younger participants, up to five years for older participants) and smoking status (never, past, current). Cases were "pair" matched 1:1, 1:2 or 1:3 with a control participant(s). Due to DNA availability samples were genotyped in two batches on the same platform at the same genotyping center at different time points.

Postmenopausal Hormones Supplementary Study to the Colon Cancer Family Registry (PMH-CCFR)

Eligible case patients included all female residents, ages 50 to 74 years, residing in the 13 counties in Washington State reporting to the Cancer Surveillance SEER program, who were newly diagnosed with invasive colorectal adenocarcinoma (ICD-O C18.0, C18.2-.9, C19.9, C20.0-.9) between October 1998 and February 2002[25]. Eligibility for all individuals was limited to those who were English-speaking with available telephone numbers, in which they could be contacted. On average, cases were identified within four months of diagnosis. The overall response proportion of eligible cases identified was 73%. Community-based controls were randomly selected according to age distribution (in 5-year age intervals) of the eligible cases by using lists of licensed drivers from the Washington State Department of Licensing for individuals, ages 50 to 64 years, and rosters from the Health Care Financing Administration (now the Centers for Medicare and Medicaid) for individuals older than 64 years. The overall response proportion of eligible controls was 66%. In GECCO, samples with sufficient DNA extracted from blood were genotyped. Only participants that were not part of the CCFR Seattle site were included in the sample set.

Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)

PLCO enrolled 154,934 participants (men and women, aged between 55 and 74 years) at ten centers into a large, randomized, two-arm trial to determine the effectiveness of screening to reduce cancer mortality. Sequential blood samples were collected from participants assigned to the screening arm. Participation was 93% at the baseline blood draw. In the observational (control) arm, buccal cells were collected via mail using the “swish-and-spit” protocol and participation rate was 65%. Details of this study have been previously described[26,27] and are

available online (http://dcp.cancer.gov/plco). The Set 1 scan included a subset of 577 colon cancer cases self-reported as being non-Hispanic White with available DNA samples, questionnaire data, and appropriate consent for ancillary epidemiologic studies. Cases were excluded if they had a history of inflammatory bowel disease, polyps, polyposis syndrome or cancer (excluding basal or squamous cell skin cancer). Controls come from the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer scan 18[28], (all male) and the GWAS of Lung Cancer and Smoking[29] (enriched for smokers) along with an additional 92 non-Hispanic White female controls. For the Set 2 scan, cases were colorectal cancers from both arms of the trial, which were not already included in Set 1. Samples were excluded if participants did not sign appropriate consents, if DNA was unavailable, if baseline questionnaire data with follow-up were unavailable, if they had a history of colon cancer prior to the trial, if they were a rare cancer, and if they were already in colon GWAS, or if they were a control in the prostate or lung populations. Controls were frequency matched 1:1 to cases without replacement, and cases were not eligible to be controls. Matching criteria were age at enrollment (two-year blocks), enrollment date (two-year blocks), sex, race / ethnicity, trial arm, and study year of diagnosis (i.e. controls must be cancer free into the case's year of diagnosis).

The Swedish Mammography Cohort (SMC) and the Cohort of Swedish Men (COSM)

SMC and COSM are two large population-based prospective cohorts from central Sweden[30]. The SMC was initiated between 1987 and 1990 when all women born in 1914-1948 and residing in Uppsala and Västmanland counties were invited; response rate 74% (n = 66,651). The COSM started in late 1997, with the invitation of all men born in 1918-1952 and residing in Västmanland and Örebro county; response rate 49% (n = 48,850). Questionnaire data on diet and other lifestyle factors was collected at the start of the studies, and has been updated repeatedly during follow-up. Further, biological samples (saliva, blood) have been collected together with

signed informed consent and are available for DNA extraction. The cohorts are annually matched to the Swedish Cancer Register for ascertainment of incident cancer cases. For the CORECT study, follow-up through 2011 was available. The Regional Ethical Review Board at Karolinska Institutet in Stockholm approved genetic studies of CRC based on the cohorts. The analytic dataset from this study included in the CORECT Phase 2 GWAS consisted of 580 CRC cases and 859 controls.

VITamins And Lifestyle (VITAL)

The VITAL cohort comprises of 77,721 Washington State men and women aged 50 to 76 years, recruited from 2000 to 2002 to investigate the association of supplement use and lifestyle factors with cancer risk[31]. Subjects were recruited by mail, from October 2000 to December 2002, using names purchased from a commercial mailing list. All subjects completed a 24-page questionnaire and buccal-cell specimens for DNA were self-collected by 70% of the participants. Subjects are followed for cancer by linkage to the western Washington SEER cancer registry and are censored when they move out of the area covered by the registry or at time of death. Details of this study have been previously described[31]. In GECCO, a nested case-control set was genotyped. Samples included, colorectal cancer cases with DNA, excluding subject with colorectal cancer before baseline, in situ cases, (large cell) neuroendocrine carcinoma, squamous cell carcinoma, carcinoid tumor, Goblet cell carcinoid, any type of lymphoma, including non-Hodgkin, Mantle cell, large B-cell, or follicular lymphoma. Controls were matched on age at enrollment (within one year), enrollment date (within one year), sex, and race / ethnicity. One control was randomly selected per case among all controls that matched on the four factors above and where the control follow-up time was greater than follow-up time of the case until diagnosis.

Women's Health Initiative (WHI)

WHI is a long-term health study of 161,808 postmenopausal women aged 50 to 79 years at 40 clinical centers throughout the U.S[32,33]. WHI comprises a Clinical Trial (CT) arm, an Observational Study (OS) arm, and several extension studies. The details of WHI have been previously described[32,33] and are available online (https://cleo.whi.org/SitePages/Home.aspx). In GECCO, Set 1 cases were selected from the September 12, 2005 database and were comprised of centrally adjudicated colon cancer cases from the Observational Study (OS) who self-reported as White. Controls were first selected among controls previously genotyped as part of a Hip Fracture GWAS conducted within the WHI OS and matched to cases on age (within three years) enrollment date (within 365 days), hysterectomy status, and prevalent conditions at baseline. For 37 cases, there was not a control match in the Hip Fracture GWAS. For these participants, we identified a matched control in the WHI OS based on same criteria. In the Set 2 scan, cases were selected from the August 2009 database and were comprised of centrally adjudicated colon and colorectal cancer cases from the OS and CT who were not genotyped in Set 1. In addition, case and control participants were subject to the following exclusion criteria: a prior history of colorectal cancer at baseline, IRB approval not available for data submission into dbGaP, and not sufficient DNA available. Matching criteria included age (within years), race/ethnicity, WHI date (within three years), WHI Calcium and Vitamin D study date (within three years), and randomization arms (OS flag, hormone therapy assignments, dietary modification assignments, calcium/vitamin D assignments). In addition, they were matched on the four regions of randomization centers. Each case was matched with one control (1:1) that exactly met the matching criteria. Control selection was done in a time-forward manner, selecting one control for each case first from the risk set at the time of the case's event. The matching algorithm allowed the closest match based on a criterion to minimize an overall distance measure[34]. Each matching

factor was given the same weight. Additional available controls that were genotyped as part of the Hip Fracture GWAS were included to improve power.

REFERENCES

1. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. *Ann Epidemiol.* 1994;4(1):1-10.

2. Alpha-Tocopherol BCCPSG. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med.* 1994;330(15):1029-1035.

3. Calle EE, Rodriguez C, Jacobs EJ, et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer.* 2002;94(9):2490-2501.

4. Kury S, Buecher B, Robiou-du-Pont S, et al. Combinations of cytochrome P450 gene polymorphisms enhancing the risk for sporadic colorectal cancer related to red meat consumption. *Cancer Epidemiol Biomarkers Prev.* 2007;16(7):1460-1467.

5. Newcomb PA, Baron J, Cotterchio M, et al. Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev.* 2007;16(11):2331-2343.

6. Figueiredo JC, Lewinger JP, Song C, et al. Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. *Cancer Epidemiol Biomarkers Prev.* 2011;20(5):758-766.

7. Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Annals of internal medicine.* 2011;154(1):22-30.

8. Lilla C, Verla-Tebit E, Risch A, et al. Effect of NAT1 and NAT2 genetic polymorphisms on colorectal cancer risk associated with exposure to tobacco smoke and meat consumption. *Cancer Epidemiol Biomarkers Prev.* 2006;15(1):99-107.

9. Slattery ML, Berry TD, Potter J, Caan B. Diet diversity, diet composition, and risk of colon cancer (United States). *Cancer Causes Control.* 1997;8(6):872-882.

10. Le Marchand L, Hankin JH, Wilkens LR, et al. Combined effects of well-done red meat, smoking, and rapid N-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev.* 2001;10(12):1259-1266.

11. Rimm EB, Stampfer MJ, Colditz GA, Chute CG, Litin LB, Willett WC. Validity of self-

reported waist and hip circumferences in men and women. *Epidemiology.* 1990;1(6):466-473.

12. Li D, Conti DV. Detecting gene-environment interactions using a combined case-only and case-control approach. *Am J Epidemiol.* 2009;169(4):497-504.

13. Helmus DS, Thompson CL, Zelenskiy S, Tucker TC, Li L. Red meat-derived heterocyclic amines increase risk of colon cancer: a population-based case-control study. *Nutr Cancer.* 2013;65(8):1141-1150.

14. Milne RL, Fletcher AS, MacInnis RJ, et al. Cohort Profile: The Melbourne Collaborative Cohort Study (Health 2020). *Int J Epidemiol.* 2017;46(6):1757-1757i.

15. Giles GG, English DR. The Melbourne Collaborative Cohort Study. *IARC Sci Publ.* 2002;156:69-70.

16. Poynter JN, Gruber SB, Higgins PD, et al. Statins and the risk of colorectal cancer. *N Engl J Med.* 2005;352(21):2184-2192.

17. Kolonel LN, Henderson BE, Hankin JH, et al. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol.* 2000;151(4):346-357.

18. Woods MO, Younghusband HB, Parfrey PS, et al. The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut.* 2010;59(10):1369-1377.

19. Belanger CF, Hennekens CH, Rosner B, Speizer FE. The nurses' health study. *Am J Nurs.* 1978;78(6):1039-1040.

20. Cotterchio M, McKeown-Eyssen G, Sutherland H, et al. Ontario familial colon cancer registry: methods and first-year response rates. *Chronic Dis Can.* 2000;21(2):81-86.

21. Cotterchio M, Manno M, Klar N, McLaughlin J, Gallinger S. Colorectal screening is associated with reduced colorectal cancer risk: a case-control study within the population-based Ontario Familial Colorectal Cancer Registry. *Cancer Causes Control.* 2005;16(7):865-875.

22. Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet.* 2007;39(8):989-994.

23. Christen WG, Gaziano JM, Hennekens CH. Design of Physicians' Health Study II--a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Ann Epidemiol.* 2000;10(2):125-134.

24. Hennekens CH, Eberlein K. A randomized trial of aspirin and beta-carotene among U.S. physicians. *Prev Med.* 1985;14(2):165-168.

25. Newcomb PA, Zheng Y, Chia VM, et al. Estrogen plus progestin use, microsatellite instability, and the risk of colorectal cancer in women. *Cancer Res.* 2007;67(15):7534-7539.

26. Gohagan JK, Prorok PC, Hayes RB, Kramer BS, Prostate LC, Ovarian Cancer Screening Trial Project T. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control Clin Trials.* 2000;21(6 Suppl):251S-272S.

27. Prorok PC, Andriole GL, Bresalier RS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin Trials.* 2000;21(6 Suppl):273S-309S.

28. Yeager M, Chatterjee N, Ciampa J, et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet.* 2009;41(10):1055-1057.

29. Landi MT, Chatterjee N, Yu K, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet.* 2009;85(5):679-691.

30. Harris H, Håkansson N, Olofsson C, et al. The Swedish mammography cohort and the cohort of Swedish men: study design and characteristics of two populationbased longitudinal cohorts. *OA Epidemiology.* 2013;1(2):16.

31. White E, Patterson RE, Kristal AR, et al. VITamins And Lifestyle cohort study: study design and characteristics of supplement users. *Am J Epidemiol.* 2004;159(1):83-93.

32. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials.* 1998;19(1):61-109.

33. Hays J, Hunt JR, Hubbell FA, et al. The Women's Health Initiative recruitment methods and results. *Ann Epidemiol.* 2003;13(9 Suppl):S18-77.

34. Bergstralh EH KJ. Computerized matching of cases to controls. *Technical Report Number 56.* 1995.

Supplemental Text 2. Description of Studies for Chapter 3

European Prospective Investigation into Cancer and Nutrition (EPIC)

EPIC is an ongoing multicenter prospective cohort study designed to investigate the associations between diet, lifestyle, genetic and environmental factors and various types of cancer[1]. In summary, 521,448 participants (~70% women) mostly aged 35 years or above were recruited between 1992 and 2000. Participants were recruited from 23 study centers in ten European countries. The current study included participants from France, Germany, Greece, Italy, the Netherlands, Spain, Sweden, and United Kingdom (UK). Blood samples were collected at baseline according to standardized procedures, and stored at the International Agency for Research on Cancer (IARC; -196°C, liquid nitrogen) for all countries except Sweden (-80°C freezers). All study participants provided written informed consent. Ethical approval for the EPIC study was obtained from the review boards of IARC and local participating centers. Incident cancer cases were identified using population cancer registries in Italy, the Netherlands, Spain, and the United Kingdom. In Sweden, cases were identified by linkage with the essentially complete Cancer Registry of Northern Sweden and were verified by a gastrointestinal pathologist. In France, Germany and Greece, cancer cases were identified during follow-up by a combination of methods including health insurance records, cancer and pathology registries, and by active follow-up directly through study participants or through next-of-kin. Controls were selected from the full cohort of individuals who were alive and free of cancer (except non-melanoma skin cancer) at the time of diagnoses of the cases, using incidence density sampling and matched by: age (±6 months at recruitment), sex, study center, follow-up time since blood collection, time of day at blood collection (±4 hours), fasting status, menopausal status, and phase of menstrual cycle at blood collection. The GECCO data set includes 2,095 incident colorectal cancer cases, and 2,306 matched controls, genotyped using the

HumanOmniExpressExome-8v1-2 array. Of these, 758 incident colorectal cancer cases, and 692 controls with available CRP data were included in this analysis.

Hawaiian Case Control Adenoma Study (HawaiiCCS)

For this adenoma study, two flexible-sigmoidoscopy screening clinics were first used to recruit participants on Oahu, Hawaii. Adenoma cases were identified either from the baseline examination at the Hawaii site of the Prostate Lung Colorectal and Ovarian cancer screening trial during 1996–2000 or at the Kaiser Permanente Hawaii's Gastroenterology Screening Clinic during 1995–2007. In addition, starting in 2002 and up to 2007, we also approached for recruitment all eligible patients who underwent a colonoscopy in the Kaiser Permanente Hawaii Gastroenterology Department. Cases were patients with histologically confirmed first-time adenoma(s) of the colorectum and were of Japanese, Caucasian or Hawaiian race/ethnicity. Controls were selected among patients with a normal colorectum and were individually matched to the cases on age at exam, sex, race/ethnicity, screening date (±3 months) and clinic and type of examination (colonoscopy or flexible sigmoidoscopy). We recruited 1016 adenoma cases (67.8% of all eligible) and 1355 controls (69.2% of all eligible); 889 cases and 1169 controls agreed to give a blood and 29 cases and 34 controls, a mouthwash sample. The GECCO data set includes a total of 989 cases and 1185 controls selected for genotyping using the OncoArray+custom iSelect array. The analyses described here only included the subset of European-ancestry individuals that also had CRP data available (49 Cases, 493 Controls).

Health Professionals Follow-up Study (HPFS)

The HPFS is a parallel prospective study to the Nurses' Health Study (NHS)[2]. The HPFS cohort comprises 51,529 men who, in 1986, responded to a mailed questionnaire. The participants are U.S. male dentists, optometrists, osteopaths, podiatrists, pharmacists, and veterinarians born

between 1910 and 1946. Participants have provided information on health-related exposures, including: current and past smoking history, age, weight, height, diet, physical activity, aspirin use, and family history of colorectal cancer. Colorectal cancer and other outcomes were reported by participants or next-of-kin and followed up through review of the medical and pathology record by physicians. Overall, more than 97% of self-reported colorectal cancers were confirmed by medical record review. Information was abstracted on histology and primary location. Incident cases are defined as those occurring after the subject provided the blood sample. Prevalent cases are defined as those occurring after enrollment in the study, but prior to the subject providing the blood sample. Follow-up has been excellent, with 94% of the men responding to date. Colorectal cancer cases were ascertained through January 1, 2008. In 1993-95, 18,825 men in HPFS mailed in blood samples by overnight courier which were aliquoted into buffy coat and stored in liquid nitrogen. In 2001-04, 13,956 men in HPFS who had not previously provided a blood sample mailed in a "swish-and-spit" sample of buccal cells. Incident cases are defined as those occurring after the subject provided a blood or buccal sample. Prevalent cases are defined as those occurring after enrollment in the study in 1986, but prior to the subject providing either a blood or buccal sample. After excluding participants with histories of cancer (except non-melanoma skin), ulcerative colitis, or familial polyposis, two case-control sets were constructed from which DNA was isolated from either buffy coat or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the cases; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the case. For both case-control

sets, matching criteria included year of birth (within 1 year) and month/year of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s). A subset of 255 cases and 258 controls from the GECCO data set that also had CRP data available were used for this analysis.

Nurses' Health Study (NHS)

The NHS cohort began in 1976 when 121,700 married female registered nurses aged 30 to 55 years returned the initial questionnaire that ascertained a variety of important health-related exposures [3]. Since 1976, follow-up questionnaires have been mailed every two years. Colorectal cancer and other outcomes were reported by participants or next-of-kin and followed up through review of the medical and pathology record by physicians. Overall, more than 97% of self-reported colorectal cancers were confirmed by medical-record review. Information was abstracted on histology and primary location. Follow-up has been high: as a proportion of the total possible follow-up time, follow-up has been over 92%. Colorectal cancer cases were ascertained through June 1, 2008. In 1989-90, 32,826 women in NHS I, mailed in blood samples by overnight courier which were aliquoted into buffy coat and stored in liquid nitrogen. In 2001-04, 29,684 women in NHS I who did not previously provide a blood sample mailed in a "swish-and-spit" sample of buccal cells. Incident cases are defined as those occurring after the subject provided a blood or buccal sample. Prevalent cases are defined as those occurring after enrollment in the study in 1976, but prior to the subject providing either a blood or buccal sample. After excluding participants with histories of cancer (except non-melanoma skin), ulcerative colitis, or familial polyposis, two case-control sets were constructed from which DNA was isolated from either buffy coat or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample

and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the case; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the cases. For both case-control sets, matching criteria included year of birth (within one year) and month / year of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s). A subset of 305 cases and 507 controls from the GECCO data set that also had CRP data available were used for this analysis.

Women's Health Initiative (WHI)

WHI is a long-term health study of 161,808 postmenopausal women aged 50 to 79 years at 40 clinical centers throughout the U.S [4,5]. WHI comprises a Clinical Trial (CT) arm, an Observational Study (OS) arm, and several extension studies. The details of WHI have been previously described [4,5] and are available online (https://cleo.whi.org/SitePages/Home.aspx). In GECCO, Set 1 cases were selected from the September 12, 2005 database and were comprised of centrally adjudicated colon cancer cases from the Observational Study (OS) who self-reported as White. Controls were first selected among controls previously genotyped as part of a Hip Fracture GWAS conducted within the WHI OS and matched to cases on age (within three years) enrollment date (within 365 days), hysterectomy status, and prevalent conditions at baseline. For 37 cases, there was not a control match in the Hip Fracture GWAS. For these participants, we identified a matched control in the WHI OS based on same criteria. In the Set 2 scan, cases were selected from the August 2009 database and were comprised of centrally adjudicated colon and colorectal cancer cases from the OS and CT who were not genotyped in Set 1. In addition, case and control participants were subject to the following exclusion criteria: a prior history of

colorectal cancer at baseline, IRB approval not available for data submission into dbGaP, and not sufficient DNA available. Matching criteria included age (within years), race/ethnicity, WHI date (within three years), WHI Calcium and Vitamin D study date (within three years), and randomization arms (OS flag, hormone therapy assignments, dietary modification assignments, calcium/vitamin D assignments). In addition, they were matched on the four regions of randomization centers. Each case was matched with one control (1:1) that exactly met the matching criteria. Control selection was done in a time-forward manner, selecting one control for each case first from the risk set at the time of the case's event. The matching algorithm allowed the closest match based on a criterion to minimize an overall distance measure [6]. Each matching factor was given the same weight. Additional available controls that were genotyped as part of the Hip Fracture GWAS were included to improve power. In total, 1268 cases and 852 controls from the GECCO data set that also had CRP data available were used for this analysis.

REFERENCES

1. Riboli E. The European Prospective Investigation into Cancer and Nutrition (EPIC): plans and progress. *J Nutr.* 2001;131(1):170S-175S.

2. Rimm EB, Stampfer MJ, Colditz GA, Chute CG, Litin LB, Willett WC. Validity of self-reported waist and hip circumferences in men and women. *Epidemiology.* 1990;1(6):466-473.

3. Belanger CF, Hennekens CH, Rosner B, Speizer FE. The nurses' health study. *Am J Nurs.* 1978;78(6):1039-1040.

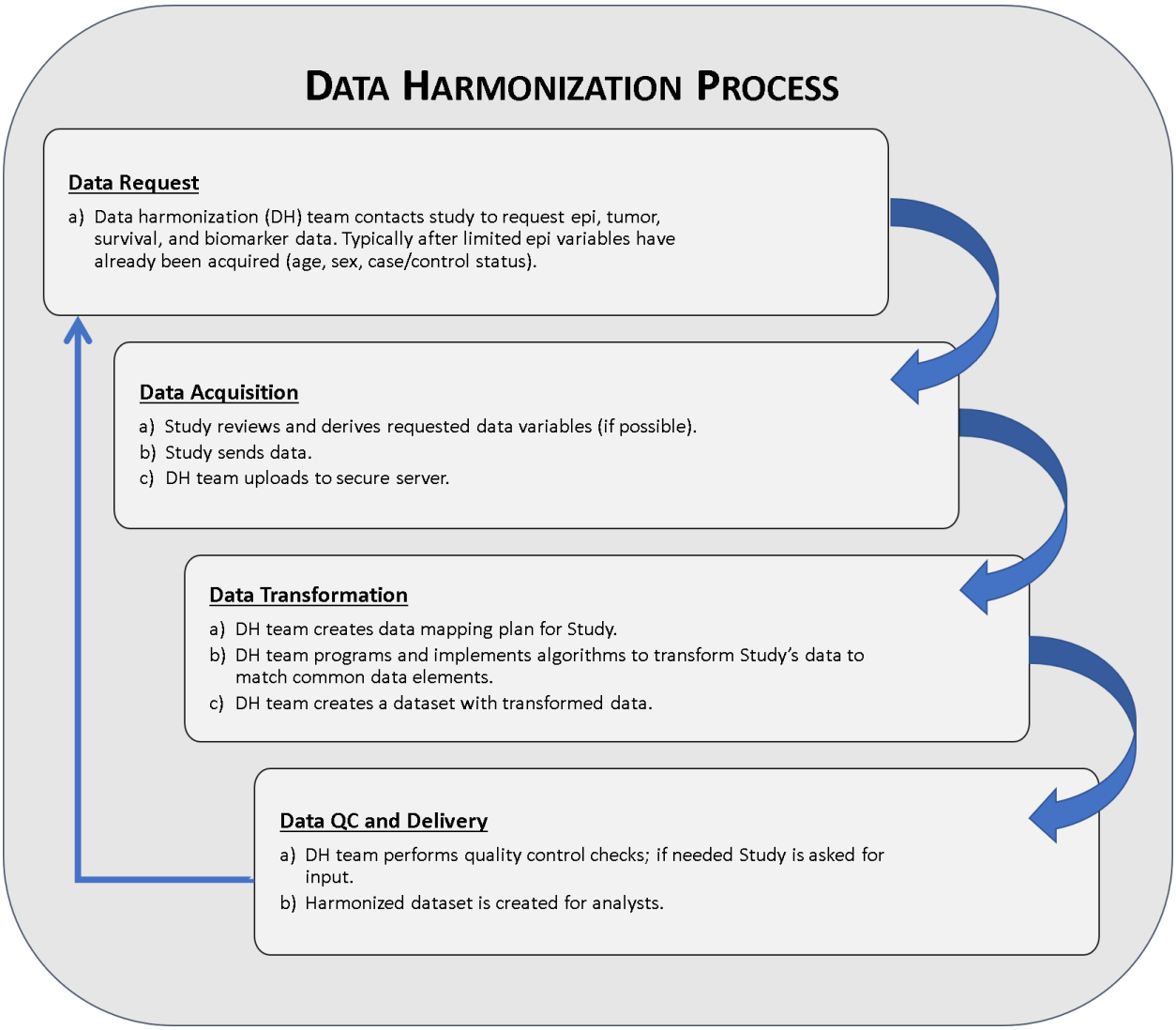
4. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials.* 1998;19(1):61-109.

5. Hays J, Hunt JR, Hubbell FA, et al. The Women's Health Initiative recruitment methods and results. *Ann Epidemiol.* 2003;13(9 Suppl):S18-77.

6. Bergstralh EH KJ. Computerized matching of cases to controls. *Technical Report Number 56.* 1995.

APPENDIX B: SUPPLEMENTAL FIGURES

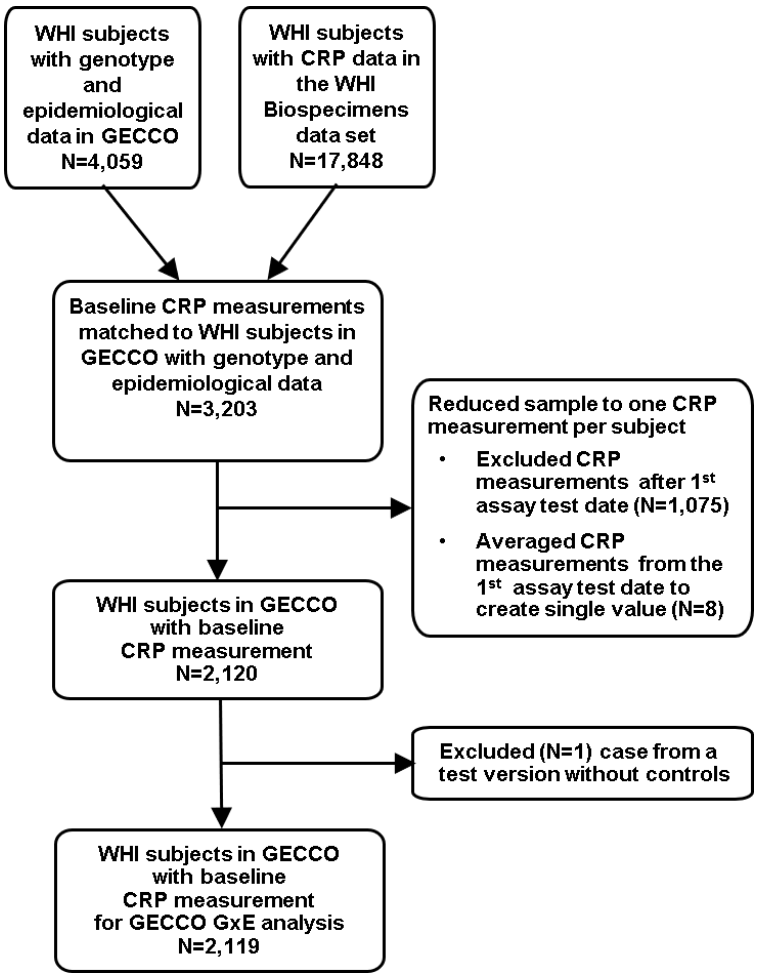
Supplemental Figure 1. Flowchart of the data harmonization process (Chapter 2 & 3)



The data harmonization process began by defining common data elements (CDEs). We then examined questionnaires and data dictionaries for each study to identify study specific data elements that could be mapped to the CDEs. CDEs were then requested from studies. Through an iterative process, we communicated with each data contributor to obtain relevant data documentation and variable coding information. The data elements were written to a common data platform, transformed, and combined into a single dataset with common definitions, standardized permissible

values, and standardized coding. The mapping and resulting data were reviewed for quality assurance, using range and logic checks to assess data and to examine intra- and inter-study variable distributions. Variables with outlying values were truncated to the minimum or maximum value of established ranges for each variable. After QC, a harmonized data set was delivered to a secured platform for use by analysts

Supplemental Figure 2: Selection of CRP measurements from WHI sub-studies for analysis



Here we aim to clarify the process by which single measurements were selected for WHI subjects for this analysis. First, we matched WHI subjects with genotype and epidemiological data in GECCO (N=4,059) to all WHI subjects with baseline CRP measurements in the WHI Biospecimen data set (N=17,848). There were multiple CRP measurements for some WHI subjects because of intra-study overlap or repeat longitudinal measures. WHI subjects with genotype and epidemiological data in GECCO had a total of 3,202 CRP baseline measurements available for analysis. We sorted WHI CRP measurements for subjects from baseline blood

draws by ascending dates of laboratory assay test. We selected only the CRP measurement from the earliest laboratory test date. For subjects with more than one CRP measurement on the earliest laboratory test date (N=8), we averaged the CRP measurements to create a single baseline CRP measurement. We then examined the 2,120 CRP measurements by test version. WHI assigns a new test version ID for an analyte for every new test method, sample type used, unit of measure used, or laboratory that tests the analyte. The 2,120 CRP measurements from WHI spanned fourteen test versions from fourteen sub-studies (main or ancillary studies in WHI). We excluded data from one test version because it did not have adequate balance between cases and controls (N=1). The final sample was comprised of 2,119 subjects with CRP measurements that we then examined stratified by WHI sub-study and outcome. We used factorial ANOVA tests to ensure there was no significant variation amongst cases or controls within our final WHI sample.

APPENDIX C: SUPPLEMENTAL TABLES

Supplemental Table 1. Genotyping platforms for participating studies (Chapter 2 & 3)

	Study Name	Acronym	Genotyping Platform	N	Cases	Controls
Data Set 1	Association STudy Evaluating RISK for sporadic colorectal cancer	ASTERISK	Illumina 300K	1820	877	943
	Colon Cancer Family Registry	CCFR_1	Illumina 1M, 1M duo	2000	1027	973
	Colon Cancer Family Registry	CCFR_2	Illumina 1M, 1M duo	443	208	235
	Colon Cancer Family Registry	CCFR_3	Affymetrix Axiom	1651	876	775
	Colon Cancer Family Registry	CCFR_4	Illumina Octorara	1674	1127	547
	Hawai'i Colorectal Cancer Studies 2&3	Colo2&3	Illumina 300K	208	87	121
	Darmkrebs: Chancen der Verhütung durch Screening	DACHS_1	Illumina Oncoarray	449	220	229
	Darmkrebs: Chancen der Verhütung durch Screening	DACHS_2	Affymetrix Axiom	398	276	122
	Diet, Activity and Lifestyle Study	DALS_1	Illumina 550K, 610K	2274	1,104	1,170
	Health Professionals Follow-Up Study	HPFS_1	Illumina OmniExpress	401	173	228
	Health Professionals Follow-Up Adenoma Study	HPFS_AD	Illumina OmniExpress	655	312	343
	Multiethnic Cohort Study	MEC_1	Illumina 300K	656	312	344
	Nurses' Health Study	NHS_1	Illumina OmniExpress	1071	297	774
	Nurses' Health Adenoma Study	NHS_AD	Illumina OmniExpress	1090	513	577
	Ontario Familial Colorectal Cancer Registry	OFCCR	Affymetrix 500K	1016	498	518
	Physicians' Health Study	PHS	Illumina OmniExpress	762	373	389
	Postmenopausal Hormones Supplementary Study to the Colon Cancer Family Registry	PMH-CCFR	Illumina 300K	398	276	122
	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	PLCO_1	Illumina 300/240S & 610K	612	201	411
	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	PLCO_2	Illumina 300K	449	220	229
	VITamins And Lifestyle	VITAL	Illumina 300K	511	249	262
	Women's Health Initiative Study	WHI_1	Illumina 550K, 550Kduo, 610K	975	455	520
	Women's Health Initiative Study	WHI_2	Illumina 300K	1982	976	1006
Data Set 2	Alpha-Tocopherol, Beta Carotene Cancer Prevention Study	ATBC	Illumina Oncoarray	511	249	229
	American Cancer Society Cancer Prevention Study II	CPSII_1	Affymetrix Axiom	975	455	229
	Colorectal Cancer Genetics & Genomics, Spanish study	CRCGEN	Affymetrix Axiom	1982	976	229
	Kentucky Case-Control Study	Kentucky	Illumina OmniExpress	3407	1705	229
	Melbourne Collaborative Cohort Study	MCCS	Affymetrix Axiom	1164	666	122
	Molecular Epidemiology of Colorectal Cancer Study	MECC_1	Illumina Oncoarray	165	137	262
	Molecular Epidemiology of Colorectal Cancer Study	MECC_2	Affymetrix Axiom	988	490	520
	Molecular Epidemiology of Colorectal Cancer Study	MECC_3	Illumina Oncoarray	1532	705	1006
	Newfoundland Case-Control Study	NFCCR	Affymetrix Axiom	641	184	457
	Swedish Mammography Cohort and Cohort of Swedish Men	SMC_COSM	Illumina Oncoarray	914	419	495

Supplemental Table 2. Missingness of dietary factor by outcome in GECCO

Study	N		Red meat NAs		Processed meat NAs		Vegetable NAs		Fruit NAs		Fiber NAs	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
ASTERISK	877	943	1.9%	0.5%	100.0%	100.0%	0.9%	0.6%	1.3%	0.8%	100.0%	100.0%
ATBC	137	28	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
CCFR	3,238	2,530	12.2%	7.6%	71.4%	65.1%	0.8%	0.8%	2.9%	2.2%	71.4%	65.1%
Colo23	87	121	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
CPSII	490	498	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
CRCGEN	705	827	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	1.7%
DACHS	2,371	2,200	0.0%	0.4%	0.0%	0.4%	0.2%	0.1%	0.4%	0.1%	100.0%	100.0%
DALS	1,104	1,170	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
HPFS	485	571	0.0%	0.0%	3.7%	4.2%	2.5%	3.0%	2.5%	3.0%	4.7%	4.7%
Kentucky	929	1,026	1.2%	1.7%	1.2%	1.9%	0.3%	0.2%	1.2%	0.9%	0.0%	0.0%
MCCS	752	671	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
MEC	312	334	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
MECC	4,087	3,159	0.4%	0.0%	0.4%	0.0%	0.1%	0.0%	0.1%	0.1%	0.0%	0.0%
NFCCR	184	457	2.2%	1.1%	100.0%	100.0%	2.7%	1.8%	8.7%	5.5%	0.0%	0.0%
NHS	811	1,351	0.0%	0.0%	4.1%	4.2%	3.3%	4.0%	3.3%	4.0%	4.4%	4.6%
OFCCR	498	518	4.4%	4.1%	100.0%	100.0%	0.8%	0.4%	3.6%	3.1%	100.0%	100.0%
PHS	373	389	1.6%	2.3%	1.6%	2.3%	0.8%	0.3%	0.8%	0.3%	100.0%	100.0%
PLCO	421	640	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
PMH-CCFR	276	122	4.0%	7.4%	100.0%	100.0%	0.4%	0.0%	2.5%	0.0%	100.0%	100.0%
SMC_COSM	419	495	1.0%	0.6%	2.1%	1.6%	0.5%	0.4%	1.4%	0.8%	0.0%	0.0%
VITAL	249	262	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
WHI	1,431	1,526	0.1%	0.2%	0.0%	0.0%	0.1%	0.2%	0.1%	0.2%	0.1%	0.2%
Overall	**20,236**	**19,838**	**2.4%**	**1.4%**	**21.0%**	**19.2%**	**0.5%**	**0.6%**	**1.1%**	**1.0%**	**33.5%**	**29.9%**

Supplemental Table 3. Characteristics of subjects in GECCO with CRP data by study

	EPIC		HawaiiCCS		HPFS	
	Case (N=758)	**Control (N=691)**	**Case (N=42)**	**Control (N=493)**	**Case (N=252)**	**Control (N=258)**
CRP (mg/L)						
Mean (SD)	4.5 (6.3)	3.4 (4.7)	2.7 (3.3)	2.1 (4.0)	2.3 (3.9)	2.1 (4.0)
Median [Min, Max]	2.8 [0.2, 66.6]	2.2 [0.2, 77.9]	1.2 [0.1, 12.0]	0.8 [0.1, 48.9]	1.4 [0.1, 42.7]	1.1 [0.0, 44.6]
Sex						
Female	424 (55.9%)	366 (53.0%)	19 (45.2%)	154 (31.2%)	-	-
Menopausal status	(N=424)	(N=366)				
Pre	47 (11.1%)	40 (10.9%)	-	-	-	-
Peri	48 (11.3%)	37 (10.1%)	-	-	-	-
Post	329 (77.6%)	289 (79.0%)	-	-	-	-
PM-HRT	(N=329)	(N=289)				
Yes	86 (26.1%)	67 (23.2%)	-	-	-	-
Missing	11 (3.3%)	10 (3.5%)	-	-	-	-
Age (years)						
Mean (SD)	58.2 (8.1)	58.2 (7.8)	62.9 (6.0)	64.5 (6.9)	67.6 (8.9)	67.4 (8.9)
BMI (kg/m^2)						
Mean (SD)	27.1 (4.5)	26.5 (3.7)	28.1 (5.8)	26.7 (4.7)	26.4 (3.5)	25.4 (2.9)
Exercise						
Sedentary	435 (57.4%)	403 (58.3%)	-	-	-	-
Non-sedentary	162 (21.4%)	153 (22.1%)	-	-	-	-
Missing	161 (21.2%)	135 (19.5%)	-	-	-	-
Diabetes						
Yes	47 (6.2%)	29 (4.2%)	-	-	22 (8.7%)	7 (2.7%)
Missing	10 (1.3%)	15 (2.2%)	-	-	0 (0%)	0 (0%)
NSAID use						
Yes	-	-	-	-	120 (47.6%)	119 (46.1%)
Smoking status						
Former smoker	244 (32.2%)	217 (31.4%)	19 (45.2%)	229 (46.5%)	134 (53.2%)	121 (46.9%)
Never smoker	315 (41.6%)	298 (43.1%)	16 (38.1%)	213 (43.2%)	102 (40.5%)	116 (45.0%)
Smoker	117 (15.4%)	112 (16.2%)	7 (16.7%)	51 (10.3%)	9 (3.6%)	12 (4.7%)
Missing	82 (10.8%)	64 (9.3%)	0 (0%)	0 (0%)	7 (2.8%)	9 (3.5%)
Alcohol use						
nondrinker	222 (29.3%)	189 (27.4%)	-	-	69 (27.4%)	61 (23.6%)
1-28g/day	407 (53.7%)	417 (60.3%)	-	-	141 (56.0%)	157 (60.9%)
>28g/day	129 (17.0%)	85 (12.3%)	-	-	33 (13.1%)	34 (13.2%)
Missing	-	-	-	-	9 (3.6%)	6 (2.3%)

CRP: C-reactive protein; NSAID use: Regular use of aspirin and/or other non-steroid anti-inflammatory drug; PM-HRT: Use of post-menopausal hormone replacement therapy; SD: Standard deviation.

Supplemental Table 3 cont: Characteristics of subjects in GECCO with CRP data by study

	NHS		WHI	
	Case N=305	**Control N=505**	**Case N=1268**	**Control N=851**
CRP (mg/L)				
Mean (SD)	2.6 (3.5)	3.0 (4.3)	4.86 (6.7)	4.4 (6.3)
Median [Min, Max]	1.5 [0.1, 27.6]	1.7 [0.0, 38.0]	2.7 [0.2, 83.7]	2.5 [0.0, 76.1]
Menopausal status	N=305	N=505	N=1268	N=851
Pre	27 (8.9%)	55 (10.9%)	-	-
Peri	-	-	-	-
Post	276 (90.5%)	449 (88.9%)	1268 (100%)	851 (100%)
Missing	2 (0.7%)	1 (0.2%)	-	-
PM-HRT	N=276	N=449	N=1268	N=851
Yes	146 (52.9%)	244 (54.3%)	419 (33.0%)	417 (49.0%)
Missing	22 (8.0%)	23 (5.1%)	1 (0.1%)	-
Age (years)				
Mean (SD)	61.3 (7.7)	60.4 (7.4)	67.0 (6.5)	67.4 (6.2)
BMI (kg/m^2)				
Mean (SD)	25.9 (4.82)	25.3 (4.42)	28.3 (5.72)	27.5 (5.46)
Exercise				
Sedentary	-	-	791 (62.4%)	509 (59.8%)
Non-sedentary	-	-	427 (33.7%)	316 (37.1%)
Missing	-	-	50 (3.9%)	26 (3.1%)
Diabetes				
Yes	30 (9.8%)	16 (3.2%)	76 (6.0%)	30 (3.5%)
Missing	0 (0%)	0 (0%)	2 (0.2%)	0 (0%)
NSAID use				
Yes	110 (36.1%)	198 (39.2%)	414 (32.6%)	317 (37.3%)
Missing	0 (0%)	2 (0.4%)	21 (1.7%)	7 (0.8%)
Smoking status				
Former smoker	147 (48.2%)	231 (45.7%)	568 (44.8%)	365 (42.9%)
Never smoker	127 (41.6%)	216 (42.8%)	586 (46.2%)	425 (49.9%)
Smoker	29 (9.5%)	56 (11.1%)	91 (7.2%)	53 (6.2%)
Missing	2 (0.7%)	2 (0.4%)	23 (1.8%)	8 (0.9%)
Alcohol use				
nondrinker	128 (42.0%)	228 (45.1%)	709 (55.9%)	441 (51.8%)
1-28g/day	149 (48.9%)	224 (44.4%)	488 (38.5%)	359 (42.2%)
>28g/day	18 (5.9%)	23 (4.6%)	64 (5.0%)	48 (5.6%)
Missing	10 (3.3%)	30 (5.9%)	7 (0.6%)	3 (0.4%)

CRP: C-reactive protein; NSAID use: Regular use of aspirin and/or other non-steroid anti-inflammatory drug; PM-HRT: Use of post-menopausal hormone replacement therapy; SD: Standard deviation.

Supplemental Table 4. WHI studies with CRP measurements matched to subjects in GECCO

Study ID[a]	Study Name[b]	Specimen[c]	CV%[d]	% of WHI[e] Cases	% of WHI[e] Controls
AS83	Thrombotic, inflammatory and genetic markers for coronary heart disease in postmenopausal women: a WHI umbrella study[1]	EDTA	3.3%	15.5%	42.6%
AS105	Carotenoids in age-related eye disease study[2]	Serum	7.4%	1.1%	2.5%
AS126	Stroke risk factors and molecular markers in postmenopausal women[3]	EDTA	2.4 - 2.9%	10.3%	19.2%
AS132	A prospective study of genetic and biochemical predictors of type 2 diabetes mellitus[4]	EDTA	3.3%	1.0%	2.1%
AS133	Biochemical and genetic predictors of incident hypertension in white and black women[5]	EDTA	3.3%	0.3%	1.3%
AS179	Frailty in WHI: drugs, inflammatory and genetic markers[6]	EDTA	2.7%	1.4%	2.3%
AS195	Candidate pathways in colorectal carcinogenesis: one-carbon metabolism and inflammation[7]	Serum	3.4%	20.6%	1.8%
AS207	IGF and multiple myeloma[8]	EDTA	4.1%	20.8%	3.8%
AS284	Obesity-related pathways and risk of benign proliferative breast disease[9]	Serum	1.9%	0.3%	0.8%
BA06	Interaction effects of genes in the inflammatory pathway and dietary, supplement, and medication	Citrate	8.1%	1.7%	4.1%
BA10	Adipokines and risk of obesity-related disease[10]	Citrate	9.2%	0.7%	0.9%
BA11	Physical activity, obesity, inflammation and CHD in a multi-ethnic cohort of women[11]	EDTA	4.1%	0.5%	0.8%
BA20	Evaluation of specific markers of rheumatoid arthritis, Inflammation, thrombogenesis and risk of cardiovascular disease and total mortality[12]	EDTA	2.1%	0.6%	0.7%
BA21	Understanding the role of sex hormones in colorectal cancer[13]	Serum	4.0%	1.7%	3.9%
BA22	Predictive modeling for CVD in a multiethnic cohort in women[14]	EDTA	4.1%	2.8%	2.8%
W58	CVD, diabetes, and renal biomarkers in the EA HT Cohort[15]	Serum	2.3%	1.9%	2.2%
W06	HT CVD Biomarkers: study of CHD, Stroke and VTE - Phase I[16]	Citrate	9.2%	1.4%	3.6%
W66	Long Life Study-Phase III Biomarkers and GWAS[15]	Serum	2.3%	17.3%	4.5%

Abbreviations: CHD: congenital heart disease; CVD: cardiovascular disease; EA: European Americans; GWAS: genome-wide association study HT: hormone therapy; IGF: insulin-like growth factor; VTE: venous thromboembolism; WHI: Women's Health Initiative.

a. Study identification designated by WHI. b. Study name designated by WHI c. Specimen source type used for CRP assay reported by WHI. d. Coefficient of variation reported by for the CRP test version(s) used by sub-study available from WHI Specimen Results Descriptions[17] e. Percent of cases and controls from each WHI sub-study that are included in our analysis: ([# of WHI sub-study cases / 852 total WHI cases) x 100%)

REFERENCES

1. Pradhan AD, Manson JE, Rossouw JE, et al. Inflammatory biomarkers, hormone replacement therapy, and incident coronary heart disease: prospective analysis from the Women's Health Initiative observational study. *JAMA.* 2002;288(8):980-987.

2. Mares JA, LaRowe TL, Snodderly DM, et al. Predictors of optical density of lutein and zeaxanthin in retinas of older women in the Carotenoids in Age-Related Eye Disease Study, an ancillary study of the Women's Health Initiative. *Am J Clin Nutr.* 2006;84(5):1107-1122.

3. Kaplan RC, McGinn AP, Baird AE, et al. Inflammation and hemostasis biomarkers for predicting stroke in postmenopausal women: the Women's Health Initiative Observational Study. *J Stroke Cerebrovasc Dis.* 2008;17(6):344-355.

4. Liu S, Tinker L, Song Y, et al. A prospective study of inflammatory cytokines and diabetes mellitus in a multiethnic cohort of postmenopausal women. *Arch Intern Med.* 2007;167(15):1676-1685.

5. Wang L, Manson JE, Gaziano JM, et al. Plasma adiponectin and the risk of hypertension in white and black postmenopausal women. *Clin Chem.* 2012;58(10):1438-1445.

6. Gray SL, LaCroix AZ, Aragaki AK, et al. Angiotensin-converting enzyme inhibitor use and incident frailty in women aged 65 and older: prospective findings from the Women's Health Initiative Observational Study. *J Am Geriatr Soc.* 2009;57(2):297-303.

7. Toriola AT, Cheng TY, Neuhouser ML, et al. Biomarkers of inflammation are associated with colorectal cancer risk in women but are not suitable as early detection markers. *Int J Cancer.* 2013;132(11):2648-2658.

8. Birmann BM, Neuhouser ML, Rosner B, et al. Prediagnosis biomarkers of insulin-like growth factor-1, insulin, and interleukin-6 dysregulation and multiple myeloma risk in the Multiple Myeloma Cohort Consortium. *Blood.* 2012;120(25):4929-4937.

9. Rohan TE, Heo M, Choi L, et al. Body fat and breast cancer risk in postmenopausal women: a longitudinal study. *J Cancer Epidemiol.* 2013;2013:754815.

10. Rajpathak SN, Kaplan RC, Wassertheil-Smoller S, et al. Resistin, but not adiponectin and leptin, is associated with the risk of ischemic stroke among postmenopausal women: results from the Women's Health Initiative. *Stroke.* 2011;42(7):1813-1820.

11. Lee IM, Sesso HD, Ridker PM, Mouton CP, Stefanick ML, Manson JE. Physical activity and inflammation in a multiethnic cohort of women. *Med Sci Sports Exerc.* 2012;44(6):1088-1096.

12. Walitt B, Mackey R, Kuller L, et al. Predictive value of autoantibody testing for validating self-reported diagnoses of rheumatoid arthritis in the Women's Health Initiative. *Am J Epidemiol.* 2013;177(9):887-893.

13. Vilasdechanon N, Kaewchur T, Ua-Apisitwong S. Radioiodine dosimetry in advance differentiated thyroid carcinoma with poor clinical outcome: A preliminary report. *J Nucl Med.* 2012;53.

14. Cook NR, Paynter NP, Eaton CB, et al. Comparison of the Framingham and Reynolds Risk scores for global cardiovascular risk prediction in the multiethnic Women's Health Initiative. *Circulation.* 2012;125(14):1748-1756, S1741-1711.

15. Coviello AD, Haring R, Wellons M, et al. A genome-wide association meta-analysis of circulating sex hormone-binding globulin reveals multiple Loci implicated in sex steroid hormone regulation. *PLoS Genet.* 2012;8(7):e1002805.

16. Wassertheil-Smoller S, Hendrix SL, Limacher M, et al. Effect of estrogen plus progestin on stroke in postmenopausal women: the Women's Health Initiative: a randomized trial. *JAMA.* 2003;289(20):2673-2684.

17. Women's Health Initiative. Specimen Results Descriptions. https://www.whi.org/researchers/data/Specimen%20Test%20Results%20DB/testList.html . Accessed 02/04/2019, 2019.

VITA

Paneen S. Petersen was born and raised with Iñupiat Ilitqusait in Alaska. In 2000, she earned a Bachelor of Science degree in Anthropology from the University of Alaska, Anchorage. In 2011, she earned a Master of Public Health degree in Epidemiology and Biostatistics from Oregon Health & Science University. In 2020, she earned a Doctor of Philosophy in Epidemiology from the University of Washington in Seattle.