A systematic analysis of interactions between environmental risk factors and genetic variation in susceptibility to colorectal cancer

Citation for published version:
Yang, T, Li, X, Farrington, SM, Dunlop, MG, Campbell, H, Timofeeva, M & Theodoratou, E 2020, 'A systematic analysis of interactions between environmental risk factors and genetic variation in susceptibility to colorectal cancer', *Cancer Epidemiology, Biomarkers and Prevention*. https://doi.org/10.1158/1055-9965.EPI-19-1328

Digital Object Identifier (DOI):
10.1158/1055-9965.EPI-19-1328

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Cancer Epidemiology, Biomarkers and Prevention

A systematic analysis of interactions between environmental risk factors and genetic variation in susceptibility to colorectal cancer

Authors

Tian Yang[1,2], Xue Li[1], Susan M Farrington[2,3], Malcolm G Dunlop[2,3], Harry Campbell[1], Maria Timofeeva[2,3]*, Evropi Theodoratou[1,2]*

* Joint and corresponding authors in this position

Affiliations

1 Centre for Global Health, Usher Institute, The University of Edinburgh, Edinburgh EH8 9AG, UK

2 Colon Cancer Genetics Group, Cancer Research UK Edinburgh Centre, Medical Research Council Institute of Genetics & Molecular Medicine, Western General Hospital, The University of Edinburgh, Edinburgh EH4 2XU, UK

3. Medical Research Council Human Genetics Unit, Medical Research Council Institute of Genetics & Molecular Medicine, Western General Hospital, The University of Edinburgh, Edinburgh EH4 2XU, UK

Corresponding authors

Evropi Theodoratou, Usher Institute, The University of Edinburgh, Old Medical School, Teviot Place, Edinburgh EH8 9AG, UK, Tel +44(0)131 650 3210, Email address: e.theodoratou@ed.ac.uk

Maria N Timofeeva, Institute of Genetics and Molecular Medicine, The University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK, Tel +44 (0) 131 651 8631, Email address: maria.timofeeva@igmm.ed.ac.uk

Running title: Gene-environment interactions in colorectal cancer risk

Keywords: colorectal cancer, gene, environment, interaction, risk factor

Financial support

The work in Scotland was funded by a Programme Grant from Cancer Research UK C348/A18927 (MGD). It was also supported by funding for the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre. This work was also funded by a grant to MGD as Project Leader with the MRC Human Genetics Unit Centre Grant (U127527202 and U127527198 from 1/4/18). TY is supported by the China Scholarship Council. ET has a Career Development Fellowship from Cancer Research UK [grant number C31250/A22804].

Disclosure of Potential Conflicts of Interest

The authors declare no potential conflicts of interest.

Word count of abstract: 250

Word count of main text: 4,017

Total number of figures and tables: 4

Number of references: 32

Abstract

Background: The underlying etiology of colorectal cancer (CRC) includes both genetic variation and environmental exposures. The main aim of this study was to search for interaction effects between well-established environmental CRC risk factors and published common genetic variants exerting main effects on CRC risk.

Methods: We used a two-phase approach: (i) Discovery phase (2,652 incident CRC cases and 10,608 controls from UK Biobank) and (ii) Validation phase (1,656 cases and 2,497 controls from the Study of Colorectal Cancer in Scotland). Interactions with nominal P<0.05 in phase I were taken forward for validation in phase II. Furthermore, we constructed a weighted genetic risk score (GRS) of CRC risk for each individual and studied interactions between the GRS and all the environmental risk factors.

Results: Seventy of the 1,500 tested interactions were found to be nominally significant in phase I. After testing these 70 interactions in phase II, the interaction between rs11903757 (2q32.3/*NABP1*) and body mass index (BMI) was nominally significant (P=0.02) with the same direction of effects. After performing fixed-effect meta-analyses to combine the results from both phases, the rs11903757*BMI interaction was also found to be statistically significant (OR=1.26; 95% CI, 1.10-1.44; $P_{interaction}=6.03\times10^{-4}$; $P_{heterogeneity}=0.63$). No interactions involving the GRS were statistically significant in either of the two datasets.

Conclusions: Limited evidence of gene-environment interactions in CRC risk was observed. There are potential modifications of the rs11903757 effect by BMI on CRC risk.

Impact: Our findings might contribute to identifying subpopulations with different susceptibility to the effect of BMI on CRC risk.

Introduction

Worldwide, colorectal cancer (CRC) is the third most common cancer by incidence and second by mortality, with over 1.8 million new cases and 881,000 deaths in 2018 (1). The underlying etiology of CRC includes both genetic variation and environmental exposures (2). It has been suggested that the interplay between genetic variants and environmental risk factors, known as gene-environment (G×E) interaction, may also contribute to "missing heritability" of CRC risk (3). Thus, identification of G×E in CRC risk should help to explain the undiscovered heritability of CRC, provide insights into CRC etiology, and identify subpopulations with high CRC risk and potential to benefit most from early intervention for CRC.

To date, few studies have explored G×E interactions on CRC risk (2, 4), partly because assembling comprehensive datasets with both risk factor exposures and genotyping data is a major challenge. We recently evaluated the evidence across the meta-analyses of candidate gene studies and genome-wide G×E interaction analyses that investigated G×E interactions in CRC (2). Notably, moderate strength of evidence was found for some G×E interactions between several single-nucleotide polymorphisms (SNPs) and alcohol drinking, processed meat intake, estrogen plus progestogen therapy use and nonsteroidal anti-inflammatory drug (NSAID) use (2).

Two recently published genome-wide association studies (GWAS) (5, 6) have identified several new genetic variants associated with CRC risk. However, the role of G×E interactions involving these GWAS-identified common genetic variants underlying CRC susceptibility remains largely unknown. In this study, we searched for interaction effects between 100 GWAS-identified independent genetic variants (linkage disequilibrium $r^2 < 0.2$) exerting main effects on CRC risk and well-established environmental CRC risk factors, including standing height, body mass index (BMI), smoking (status and pack-years of smoking), NSAID (aspirin and others) use, hormonal replacement therapy (HRT) use, physical activity, alcohol use, and dietary intakes of processed meat, red meat, vegetables, fruit, fiber and calcium. These environmental risk factors were selected based on the results of meta-analyses and systematic reviews from the World Cancer Research Fund International/American Institute for Cancer Research Third Expert Report (7) and the subsequent Continuous Update Project Report (8). In particular, it has been reported that diet low in calcium (20.5%), alcohol use (15.2%), smoking (13.3%), BMI (8.6%) and diet low in fiber (11.6%) were the risk factors that contributed most to disability-adjusted life-year estimates of CRC at the global level in 2017 (9).

Here, we utilize a two-phase approach to test for the interactions between common genetic risk factors associated with CRC at genome-wide levels of significance and environmental risk factors supported by sufficient evidence for association with CRC risk from the reports (7, 8), including: (i) Discovery phase using UK Biobank data and (ii) Validation phase using study samples from the Study of Colorectal Cancer in Scotland (SOCCS). Furthermore, we constructed a weighted genetic risk score (GRS) of CRC risk for each individual by incorporating information of the 100 independent genetic variants and studied interactions between the GRS and all the environmental risk factors.

Materials and methods

Study population

We used individual-level data from the UK Biobank cohort and SOCCS in our analysis (**Table 1**).

Case-control study from the UK Biobank cohort

The UK Biobank is a large cohort study that has recruited more than half a million people aged 40 to 69 years throughout the UK between 2006 and 2010. Questionnaire data, physical measurements, blood and urine samples were collected at the baseline assessment of UK Biobank (10). The web-based 24-hour dietary assessment was applied to collect information on the intakes of foods and beverages consumed during the 24-hour period before the assessment (11). Data abstracted from the UK Biobank study consisted of 4,800 incident and prevalent CRC cases and 20,289 population-based controls after the process of genotyping quality control (6). Of all the CRC cases, 2,652 (55.3%) were incident and 2,119 (44.1%) were prevalent cases. However, 1,907 (90.0%) prevalent CRC cases were diagnosed more than one year before recruitment, which can be a source of bias. Therefore, we only included a total of 2,652 incident CRC cases and 10,608 controls in the discovery phase of the study (**Table 1**).

Research ethics approval for UK Biobank to collect participant data was obtained from the National Information Governance Board for Health and Social Care and the North West Multicentre Research Ethics Committee. Genotypic and phenotypic data used in this study were obtained from UK Biobank under an approved data request application (application ID: 7441).

Study of Colorectal Cancer in Scotland (SOCCS)

SOCCS is a large population-based case-control study of CRC. Details of SOCCS have been described previously (12). Briefly, SOCCS has recruited cases of adenocarcinoma of colorectum who were aged 16-79 years in Scotland (12). Population-based controls who were identified through the Community Health Index were randomly invited to participate in SOCCS (12). In this study, we included a total of 1,656 CRC cases and 2,497 controls who had available phenotype and genetic data **(Table 1**).

SOCCS received research ethics approval from the MultiCentre Research Ethics Committee for Scotland and relevant Local Research Ethics committees (12). All participants provided written informed consent (12).

Genotyping and quality control

A total of 100 GWAS-identified independent genetic variants (linkage disequilibrium $r^2 < 0.2$) were examined, those identified in two recently published GWAS studies (5, 6). For the SNPs that were located at the same locus and in linkage disequilibrium, we selected the ones that were described in Law PJ, et al. (6). For the UK Biobank genotype data, biological samples of the participants were genotyped using two closely related arrays from Affymetrix: the custom-designed Affymetrix UK BiLEVE Axiom array on an initial 50,000 participants and Affymetrix UK Biobank Axiom array on the remaining 450,000 participants. The procedure of genotyping and quality control was previously reported (13). Details of phasing and imputation were previously described by Bycroft et al., 2018 (13). In brief, prediction of un-genotyped variants was done using IMPUTE4 software with a combination of reference panels including: (i) the Haplotype Reference Consortium panel; and (ii) the merged UK10K and 1000 Genome phase 3 reference panel.

For SOCCS, samples were genotyped using Illumina HumanHap300, HumanHap240S arrays (14) and OmniExpressExome BeadChip 8v1.1, 8v1.250 or 8v1.3 (Illumina Inc., San Diego, CA) (15). Un-genotyped variants were imputed using SHAPEIT v2.837 and IMPUTE v2.3.2. We used two reference panels for imputation: the 1000 Genome reference panel, phase 1, December 2013 release and the UK10K reference panel (release April 2014). For the X chromosome, genotypes were phased and imputed as for the autosomal chromosome, with the inclusion of the "chrX" flag. X chromosome variants were coded as 0 and 2 for men, assuming complete inactivation of one allele in females and equal effect-size between males and females. Details of imputation and subsequent quality control of imputed genotypes are given elsewhere (6).

Phenotype data

Cancer cases of UK Biobank were identified through data linkage to national cancer and death registries and Hospital Episode Statistics. CRC cases in UK Biobank were defined using two different revisions of the International Classification of Diseases (ICD), ICD-10 or ICD-9 (6). Height (standing and sitting) and weight were measured during the baseline physical measurement of UK Biobank (10). Information on lifestyle factors and food intakes was gathered using a self-reported touchscreen questionnaire at recruitment. Information on daily intakes of nutrients was collected using a web-based 24-hour dietary assessment tool about four years after the baseline assessment. The 24-hour dietary assessments were performed in one-third of the UK Biobank participants and was available for 947 CRC cases and 4,160 controls in our dataset. The derivation of each environmental variable in the UK Biobank dataset is described in **Supplementary Methods.** The environmental data were harmonized between the UK Biobank dataset and the SOCCS dataset whenever possible.

The CRC cases of SOCCS were defined based on histologically confirmed adenocarcinoma of the colon or rectum (codes 153 or 154 in ICD, 9th revision or ICD10 C18, C19 or C20 codes) (15). SOCCS study participants that were recruited before 2009 had completed two questionnaires: The Lifestyle and Cancer Questionnaire and The Scottish Collaborative Group Food Frequency Questionnaire (12). The derivation of each environmental variable in the SOCCS dataset has been described previously (12)**.**

Statistical methods

The association between each genetic variant, each environmental risk factor and CRC risk was examined by using logistic regression models. Within the UK Biobank dataset, models were adjusted for age (age of CRC diagnosis for cases and age at recruitment for controls), sex and assessment center, and analyses involving genetic variants were further adjusted for the first 10 genetic principal components. Within the SOCCS dataset, models were adjusted for age (age of CRC diagnosis for cases and age at recruitment for controls) and sex. In addition, models of the analysis of dietary nutrients were further adjusted for total energy intake.

To test for the interactions, the two-phase approach was applied. Interactions with nominal *P* values < 0.05 in phase I were further tested in phase II. Case-control logistic regression analyses including G×E interaction term(s) were applied to test for the multiplicative interactions. Models were adjusted for age, sex, assessment center and the first 10 genetic principal components in phase I (the UK Biobank dataset), whereas models were adjusted for age and sex in phase II (the SOCCS dataset). In addition, models of the interaction analysis involving dietary nutrients were further adjusted for total energy intake. Furthermore, for interactions with nominal *P* values < 0.05 in phase II, we performed fixed-effect meta-analyses to combine phase I and phase II results, and obtained summary odds ratios (ORs) and 95% confidence intervals (CIs), with estimation of heterogeneity measured by Cochran Q test (I^2) and its *P* value (16).

False Discovery Rate (17) was used to account for multiple testing in phase II. *P* values unadjusted before multiple testing were termed nominal *P* values, whereas *P* values after adjustment for multiple testing were termed adjusted *P* values and were used to evaluate the statistical significance of a given interaction at the 0.05 level.

For interactions with nominal *P* values < 0.05 in phase II, we further examined (i) the main effect of the environmental risk factor stratified by the SNP, (ii) the main effect of the SNP stratified by the environmental risk factor and (iii) the combined association stratified by both the environmental risk factor and the SNP. Also, we used an extension of the Human Genome Epidemiology Network's Venice criteria (3, 18) to evaluate the strength of the evidence for the G×E interactions with nominal *P* values < 0.05 in phase II. The detailed evaluation process has been described elsewhere (2, 3).

To calculate the weighted GRS of CRC risk for each individual, a meta-analysis excluding the UK Biobank and SOCCS study samples was first performed in order to obtain unbiased regression coefficients (β-estimates) of CRC risk associated with the genetic variants. Directly genotyped SNPs were coded as 0, 1 or 2 copies of the variant allele. For imputed SNPs, we used the estimated number of copies of the count allele (the 'dosage') with values between 0-2. Both genotyped and imputed SNPs were treated as continuous variables (i.e. log-additive model). The weighted GRS was then calculated by summing up the dosages of effect alleles weighted by their effect estimates retrieved from the meta-analysis of GWAS and were Z-transformed to normalize the distributions. Models were adjusted for the same covariates as the examination of interactions between the individual SNPs and the environmental risk factors.

Analyses were conducted using R 3.4.4 (https://www.R-project.org/). Power calculations were performed using the Quanto software (19, 20). All statistical tests were two-sided.

Results

Study characteristics are presented in **Table 1**. Briefly, 2,652 incident CRC cases and 10,608 controls from the UK Biobank cohort were included in phase I, whereas 1,656 cases and 2,497 controls from SOCCS were included in phase II. The summary statistics of the environmental risk factors for these two datasets are presented in **Supplementary Table S1**. The associations between the environmental risk factors, the 100 SNPs and CRC risk are presented in **Supplementary Tables S2 and S3**, separately.

After testing 1,500 G×E in phase I, a total of 70 G×E interactions showed nominal *P* values < 0.05 (**Supplementary Table S4)**. These interactions were further tested in phase II, in which two interactions showed nominal *P* values < 0.05, including the interactions between rs11903757 (2q32.3/*NABP1*) and BMI (nominal *P* = 0.02), and rs2735940 (5p15.33/*TERT*) and smoking status (nominal *P* = 0.04) (**Table 2**). In particular, the rs11903757*BMI interaction was found with the same direction of effects. However, neither of the two interactions reached statistical significance after accounting for multiple testing based on the 70 tests performed in phase II. After performing fixed-effect meta-analyses for these two interactions, statistical significance was observed for the interaction between rs11903757 and per 10 kg/m^2 increase in BMI (OR = 1.26; 95% CI, 1.10-1.44; $P_{interaction}$ = 6.03×10^{-4}; $P_{heterogeneity}$ = 0.63) (**Table 2**). Furthermore, the rs11903757*BMI interaction was observed with statistical significance in men after performing meta-analyses in stratified subgroups according to sex (OR = 1.32; 95% CI, 1.08-1.60; $P_{interaction} = 5.53\times10^{-3}$; $P_{heterogeneity} = 0.78$) (**Table 2**).

Stratification analyses were performed for the rs11903757*BMI interaction and the rs2735940*smoking status interaction in the UK Biobank dataset and the SOCCS dataset, respectively (**Supplementary Tables S5 to S8)**. For the rs11903757*BMI interaction**,** above median BMI significantly increased CRC risk in individuals with TC genotype (OR = 1.27; 95% CI, 1.07-1.50; $P = 5.69\times10^{-3}$) and non-significantly in individuals with CC genotype (OR = 1.32; 95% CI, 0.70-2.52; $P = 0.393$) but not in those with TT genotype ($P = 0.352$) in the UK Biobank dataset when stratified by genotypes of rs11903757 (**Supplementary Table S5)**. Also, the effect of BMI on CRC risk stratified by genotypes of rs11903757 was limited to men in the UK Biobank dataset (**Supplementary Table S5)**. For the rs2735940*smoking status interaction, ever smokers (compared to non-smokers) significantly increased CRC risk in individuals with AA genotype (OR = 1.32; 95% CI, 1.10-1.57; $P = 2.77\times10^{-3}$) but not in those with AG genotype ($P = 0.060$) or GG genotype ($P = 0.972$) in the UK Biobank dataset when stratified by genotypes of rs2735940 (**Supplementary Table S7)**.

Table 3 presents the evaluation of evidence for the rs11903757*BMI interaction and the rs2735940*smoking status interaction by using an extension of the Venice criteria (3, 18). The environmental effects of BMI on CRC risk was graded as class III (suggestive) (2) (**Supplementary Table S9**). The main effect of rs11903757

(2q32.3/*NABP1*) on CRC risk was graded as strong (AAA, based on the Venice criteria (18, 21)) in a meta-analysis of 12,696 cases and 15,113 controls of European descent (OR = 1.16; 95% CI, 1.10-1.22; P = 3.71×10^{-8}; $P_{heterogeneity}$ = 0.27) (**Supplementary Table S10**). Consequently, the interaction between rs11903757 (2q32.3/*NABP1*) and BMI was given a moderate prior score (Moderate-2) and a weak overall credibility score (**Table 3**). No evidence was found for the interaction between rs2735940 (5p15.33/*TERT*) and smoking (**Table 3**).

Table 4 presents the interaction effects between the weighted GRS and the environmental risk factors on CRC risk in the UK Biobank dataset and the SOCCS dataset. The distributions of the weighted GRS among the participants in the two datasets are shown in **Supplementary Figure 1 (A) and (B)**, respectively. The OR of the GRS was 1.64 (95% CI, 1.57-1.72; $P < 2\times10^{-16}$) in the UK Biobank dataset and 1.64 (95% CI, 1.52-1.78; $P < 2\times10^{-16}$) in the SOCCS dataset, separately. No interactions involving the GRS were statistically significant in either of the two datasets (**Table 4**).

The power to detect a G×E interaction was estimated for the phase I data at a 0.05 significance level, assuming a main effect of 1.10 for log-additive SNPs (**Supplementary Figures S2 to S4)**. We only calculated the power for binary environmental variables because statistical power would be higher for continuous exposure variables (10). The prevalence of binary environmental exposures and the environmental ORs were chosen according to the dataset used in phase I (**Supplementary Tables S1 and S2)**. With a sample size of 2,652 cases and 10,608 controls in the whole dataset, we had sufficient power (80%) to detect moderate (OR > 1.30) and strong (OR > 2.00) G×E interaction effects if the SNP was at least moderately polymorphic [minor allele frequency (MAF) = 0.20]. Similarly, the analysis of HRT use was restricted to women (1,098 cases and 4,608 controls) and the analysis of dietary nutrients was limited to the participants who had taken part in the web-based 24-hour dietary assessment of UK Biobank (947 cases and 4,160 controls in our study), we therefore had sufficient power (80%) to detect strong (OR > 2.00) G×E interaction effects for a moderately polymorphic (MAF = 0.20).

Discussion

Using a two-phase approach, followed by a fixed-effect meta-analysis, we searched for G×E interaction effects between 100 published common genetic variants and 15 environmental variables. Two of the 70 G×E interactions with nominal significance in phase I showed nominal *P* values < 0.05 in phase II, including the interactions between rs11903757 (2q32.3/*NABP1*) and BMI (*P* = 0.02), and rs2735940 (5p15.33/*TERT*) and smoking status (*P* = 0.04). In particular, the rs11903757*BMI interaction was found with the same direction of effects and showed statistical significance in the meta-analysis. No statistically significant interactions were found between the weighted GRS for CRC and the environmental risk factors in either of the two datasets.

The interaction between rs11903757 (2q32.3/*NABP1*) and BMI was nominally significant with the same direction of effects in our study. The individual effects of rs11903757 (2q32.3/*NABP1*) and BMI on CRC risk have been previously explored, but the biological mechanisms behind this interaction remains unclear. Rs11903757 is an intergenic SNP at 2q32.3 with closest proximity to the gene *nucleic acid binding protein 1 (NABP1)* (44 kb centromeric) and the gene *serum deprivation response (SDPR)* (112 kb telomeric), which encodes the serum-deprivation response phosphatidylserine-binding protein (22). Also, rs11903757 is expression quantitative trait loci for *NABP1* expression in whole blood ($P = 2.1\times10^{-15}$) and non-sun exposed skin ($P = 3.2\times10^{-6}$) based on the results from the Genotype-Tissue Expression (GTEx) project (23). Previously, the SNP rs11903757 was found to be associated with CRC risk in a GWAS of European and Asian case-control studies (OR = 1.15 per risk allele; $P = 3.7\times10^{-8}$) (22). Additionally, no statistically significant associations were observed between this genotype and CRC survival in a population-based study of 5,675 patients after CRC diagnosis in Scotland (24). The *NABP1* gene binds single-stranded DNA via the oligonucleotide/oligosaccharide binding fold domain (25). Single-stranded DNA binding proteins are essential for diverse DNA processes (22). Evidence from previous biologic data also suggests that *NABP1* plays a critical role in genomic stability, which could explain the development of cancer (26). BMI was used as a proxy variable of body fatness in our analysis because it has been reported to be strongly correlated with percentage body fat according to results from laboratory methods (10). Greater body fatness, which can be measured by BMI, waist circumference and waist-to-hip ratio, has been reported as a risk factor for CRC (8).

Hutter et al., 2012 (27) and Kantor et al., 2014 (28) performed two meta-analyses of G×E interactions between a total of 26 GWAS-identified CRC risk loci and a number of environmental factors. Only the interaction between rs16892766 (8q23.3/*EIF3H*) and vegetable consumption showed statistically significance after accounting for multiple testing in the meta-analysis with a sample size of 7,016 CRC cases and 9,723 controls (OR = 1.88; 95% CI, 1.36-2.59; nominal $P_{interaction} = 1.34\times10^{-4}$; adjusted $P_{interaction} = 0.02$; $P_{heterogeneity} = 0.68$) (27). However, this interaction was not replicated in phase I of our study (OR = 1.00; 95% CI, 0.85-1.17; $P = 0.996$). Additionally, the rs11903757*BMI (per 10 kg/m^2) interaction was found with nominal statistical significance in phase II of our study and with statistical significance in the meta-analysis, was not detected by Kantor, et al. (per 10 kg/m^2 increase; OR = 1.07; 95% CI, 0.94-1.22; $P_{interaction} = 0.28$; $P_{heterogeneity} = 0.45$) (28). There may be multiple reasons behind these observations. One of the possible reasons is that Hutter et al. (27) and Kantor et al. (28) included both nested case-control studies and case-control studies in their meta-analysis, while we only used a prospective study (including 2,652 incident CRC cases and 10,608 controls) in phase I and therefore will be less affected by recall bias or reverse causality, when weight was already affected by the presence of cancer disease.

The strengths of this study are that first, we examined for the first time the presence of potential effect-modifications for the SNPs newly identified from the two recently published meta-analyses of GWAS (5, 6). Second, we evaluated G×E interactions for a wide range of environmental CRC risk factors, for which valid information was collected across the studies. For the dataset used in phase I, we only used incident CRC cases and controls from the UK Biobank cohort. Therefore, the information on lifestyle factors and dietary habits was collected before cancer diagnosis, which minimized recall bias and differential misclassifications. Though SOCCS is a case-control study, participants were asked to provide information about general lifestyle and the consumption of each food item one year prior to diagnosis for cases and one year prior to recruitment for controls (12). Third, we critically evaluated the cumulative evidence for the identified interactions using predetermined guidelines (3, 18), which have been used to assess the cumulative evidence of G×E interaction effects on cancer risk (2, 4, 29). Lastly, for the first time we examined the interaction effects between the weighted GRS for CRC and a wide variety of environmental CRC risk factors. One study has examined joint effects between GRSs and plasma 25-hydroxyvitamin D (25(OH)D) on CRC risk (30). However, no evidence for the modification of genetic susceptibility for CRC according to vitamin D status was observed (30).

However, there are several limitations. First, our study had limited power to detect weak (OR < 1.30) or moderate (1.30 < OR < 2.00) interactions for SNPs with MAF less than 0.20 even if we used the whole dataset in phase I. Furthermore, the analysis of HRT use was restricted to women [hence has a reduced sample size and power] and information on dietary nutrient intakes in the UK Biobank cohort was collected from the web-based 24-hour dietary recall assessments [found in only one-third of all UK Biobank participant which again restricted sample size and power]. Therefore, further studies with larger sample sizes are needed to examine the interaction effects. Second, we used a prospective cohort study in phase I and a case-control study in phase II. Both types of these studies have different sources of error. For case-control studies, recall bias and differential misclassifications can bias estimates and may lead to false negatives. For our study, cases may recall their exposure better than controls because information on general lifestyle and food intakes was collected using self-reported questionnaires in SOCCS. Prospective studies can minimize differential misclassification because the information of lifestyle factors and dietary habits was collected for all participants at recruitment. However, they may have variable time period between baseline data collection and cancer diagnosis (28). In addition, we attempted to harmonize the environmental variables in the UK Biobank cohort and the SOCCS, though the two studies used different methods for data collection. Despite these concerns, the associations between the environmental risk factors and CRC risk in the prospective dataset of UK Biobank were consistent with previous observations (7, 8). Third, there is a "healthy volunteer" selection bias in UK Biobank, which means that the participants of UK Biobank are probably more aware of health issues than non-participants (31). Therefore, the UK Biobank cohort is not fully representative of the UK general population (31).

Conclusion

In conclusion, using a two-phase approach, we were able to observe a statistically significant G×E interaction between rs11903757 (2q32.3/*NABP1*) and BMI in CRC risk. Functional studies and further replications are needed to confirm our findings and uncover the mechanisms of the interactions between BMI and genetic variants. Also, larger studies incorporating information from consortia are needed to fully examine the impact of genetic variation on the effect of BMI on CRC risk, thus to provide insights into CRC etiology, and identity subpopulations who will benefit most from early intervention for CRC.

Acknowledgments

Acknowledgments for UK Biobank: This research has been conducted using the UK Biobank resource (approval number: 7441). We wish to kindly thank all participants from the UK Biobank.

Acknowledgments for the Scottish Colorectal Cancer Study: We thank the participants in all of the studies that contributed to this piece of work and all the recruitment teams and collaborators who make such studies possible. We acknowledge the excellent technical support from Marion Walker. We are grateful to Ruth Wilson, Donna Markie and all those who continue to contribute to recruitment, data collection and data curation for the Study of Colorectal Cancer in Scotland studies. In addition to all consultant colorectal surgeons who provided data on their patients, we are also indebted to the clinical networks throughout Scotland who contributed substantially to clinic pathologic data and staging information. We acknowledge the expert support on sample preparation from the Genetics Core of the Edinburgh Wellcome Trust Clinical Research Facility.

Acknowledgments for the GTEx consortium: We thank the provision of public data from the GTEx consortium. The data used for the analyses described in this article were

obtained from the GTEx Portal on 12/September/2019.

Authors' contributions

Study design: MT and ET.

Study concept: SMF, MGD and HC.

Data analysis: TY, XL and MT.

Manuscript draft and revision: TY, MT, ET, XL, SMF, MGD and HC.

Article guarantor: Dr. Maria Timofeeva and Prof. Evropi Theodoratou.

Ethical approval and consent to participate

Ethics approval for UK Biobank to collect participant data was obtained from the National Information Governance Board for Health and Social Care and the North West Multicentre Research Ethics Committee. Genotypic and phenotypic data used in this study were obtained from UK Biobank under an approved data request application (application ID: 7441). SOCCS received research ethics approval from the MultiCentre Research Ethics Committee for Scotland and relevant Local Research Ethics committees. All participants provided written informed consent.

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;68:394-424.

2. Yang T, Li X, Montazeri Z, Little J, Farrington SM, Ioannidis JP, et al. Gene-environment interactions and colorectal cancer risk: an umbrella review of systematic

reviews and meta-analyses of observational studies. Int J Cancer. 2018.

3. Boffetta P, Winn DM, Ioannidis JP, Thomas DC, Little J, Smith GD, et al. Recommendations and proposed guidelines for assessing the cumulative evidence on joint effects of genes and environments on cancer occurrence in humans. International journal of epidemiology. 2012;41:686-704.

4. Theodoratou E, Timofeeva M, Li X, Meng X, Ioannidis JPA. Nature, Nurture, and Cancer Risks: Genetic and Nutritional Contributions to Cancer. Annual review of nutrition. 2017;37:293-320.

5. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, et al. Discovery of common and rare genetic risk variants for colorectal cancer. Nature genetics. 2018.

6. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd JB, Fernandez-Tajes J, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nature Communications. 2019.

7. World Cancer Research Fund/American Institute for Cancer Research. Diet, Nutrition, Physical Activity and Cancer: a Global Perspective 2018; Available from: https://www.wcrf.org/dietandcancer

8. World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project Report: Diet, Nutrition, Physical Activity and Colorectal Cancer. 2017; Available from: https://www.wcrf.org/sites/default/files/Colorectal-Cancer-2017-Report.pdf

9. Safiri S, Sepanlou SG, Ikuta KS, Bisignano C, Salimzadeh H, Delavari A, et al. The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990&#x2013;2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet Gastroenterology & Hepatology.

10. UK Biobank. Protocol for a large-scale prospective epidemiological resource. 2007 [cited 2019; Available from: http://www.ukbiobank.ac.uk/key-documents/

11. UK Biobank. 24-hour dietary recall questionnaire (version 1.1). 2012 [cited 2019; Available from: http://www.ukbiobank.ac.uk/key-documents/

12. Theodoratou E, Farrington SM, Tenesa A, McNeill G, Cetnarskyj R, Barnetson RA, et al. Dietary vitamin B6 intake and the risk of colorectal cancer. Cancer Epidemiol Biomarkers Prev. 2008;17:171-82.

13. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203-9.

14. Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nature genetics. 2008;40:1426-35.

15. Timofeeva MN, Kinnersley B, Farrington SM, Whiffin N, Palles C, Svinti V, et al. Recurrent Coding Sequence Variation Explains Only A Small Fraction of the Genetic Architecture of Colorectal Cancer. Scientific reports. 2015;5:16286.

16. Cochran W. THE COMBINATION OF ESTIMATES FROM DIFFERENT EXPERIMENTS. Biometrics. 1954;10:101-27.

17. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and

Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57:289-300.

18. Ioannidis JP, Boffetta P, Little J, O'Brien TR, Uitterlinden AG, Vineis P, et al. Assessment of cumulative evidence on genetic associations: interim guidelines. International journal of epidemiology. 2008;37:120-32.

19. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol. 2002;155:478-84.

20. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Statistics in medicine. 2002;21:35-50.

21. Khoury MJ, Bertram L, Boffetta P, Butterworth AS, Chanock SJ, Dolan SM, et al. Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. Am J Epidemiol. 2009;170:269-79.

22. Peters U, Jiao S, Schumacher FR, Hutter CM, Aragaki AK, Baron JA, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastroenterology. 2013;144:799-807.e24.

23. GT C. The Genotype-Tissue Expression (GTEx) project. Nature genetics. 2013;45:580-5.

24. He Y, Theodoratou E, Li X, Din F, Vaughan-shaw P, Svinti MacLeod V, et al. Effects of common genetic variants associated with colorectal cancer risk on survival outcomes after diagnosis: a large population-based cohort study. International Journal of Cancer. 2019.

25. Richard DJ, Bolderson E, Cubeddu L, Wadsworth RI, Savage K, Sharma GG, et al. Single-stranded DNA-binding protein hSSB1 is critical for genomic stability. Nature. 2008;453:677-81.

26. Broderick S, Rehmet K, Concannon C, Nasheuer HP. Eukaryotic single-stranded DNA binding proteins: central factors in genome stability. Sub-cellular biochemistry. 2010;50:143-63.

27. Hutter CM, Chang-Claude J, Slattery ML, Pflugeisen BM, Lin Y, Duggan D, et al. Characterization of gene-environment interactions for colorectal cancer susceptibility loci. Cancer Res. 2012;72:2036-44.

28. Kantor ED, Hutter CM, Minnier J, Berndt SI, Brenner H, Caan BJ, et al. Gene-environment interaction involving recently identified colorectal cancer susceptibility Loci. Cancer Epidemiol Biomarkers Prev. 2014;23:1824-33.

29. Dimitrakopoulou VI, Travis RC, Shui IM, Mondul A, Albanes D, Virtamo J, et al. Interactions Between Genome-Wide Significant Genetic Variants and Circulating Concentrations of 25-Hydroxyvitamin D in Relation to Prostate Cancer Risk in the National Cancer Institute BPC3. Am J Epidemiol. 2017;185:452-64.

30. Hiraki LT, Joshi AD, Ng K, Fuchs CS, Ma J, Hazra A, et al. Joint effects of colorectal cancer susceptibility loci, circulating 25-hydroxyvitamin D and risk of colorectal cancer. PLoS ONE [Electronic Resource]. 2014;9:e92212.

31. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol. 2017;186:1026-34.

1