HHS Public Access

Author manuscript

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2022 November 04.

Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2022 May 04; 31(5): 1077–1089.

doi:10.1158/1055-9965.EPI-21-1003.

Beyond GWAS of colorectal cancer: Evidence of interaction with alcohol consumption and putative causal variant for the 10q24.2 region

A full list of authors and affiliations appears at the end of the article.

Abstract

Background: Currently known associations between common genetic variants and colorectal cancer (CRC) explain less than half of its heritability of 25%. As alcohol consumption has a J-shape association with CRC risk, non-drinking and heavy drinking are both risk factors for CRC.

Methods: Individual-level data was pooled from Colon Cancer Family Registry, Colorectal Transdisciplinary Study, and Genetics and Epidemiology of Colorectal Cancer Consortium to compare non-drinkers (<=1 g/day) and heavy drinkers (>28 g/day) with light-to-moderate drinkers (1-28 g/day) in GxE analyses. To improve power, we implemented joint 2df and 3df tests and a novel two-step method that modifies the weighted hypothesis testing framework. We prioritized putative causal variants by predicting allelic effects using support vector machine models.

Results: For non-drinking as compared to light-to-moderate drinking, the hybrid 2-step approach identified 13 significant SNPs with pairwise r2>0.9 in the 10q24.2/COX15 region. When stratified by alcohol intake, the A allele of lead SNP rs2300985 has a dose-response increase in risk of CRC as compared to the G allele in light-to-moderate drinkers (odds ratio (OR) for GA genotype=1.11, 95% confidence interval (CI)=1.06-1.17; OR for AA genotype=1.22, 95% CI=1.14-1.31), but not in non-drinkers or heavy drinkers. Among the correlated candidate SNPs in the 10q24.2/COX15 region, rs1318920 was predicted to disrupt a HNF4 transcription factor binding motif.

Conclusions: Our study suggests that the association with CRC in 10q24.2/COX15 observed in GWAS is strongest in non-drinkers. We also identified rs1318920 as the putative causal regulatory variant for the region.

Impact: The study identifies multifaceted evidence of a possible functional effect for rs1318920

Introduction

Though alcohol consumption is considered a risk factor for colorectal cancer (CRC), meta-analyses across our large consortia have revealed a J-shape relationship with alcohol consumption. Light-to-moderate drinking is the group at the lowest risk of CRC, while the risk of CRC increases slightly in non-drinkers and substantially in very heavy drinkers [1]. Many mechanisms have been proposed to explain the relationship between alcohol consumption and colon carcinogenesis [2], but the lower risk of CRC observed among

A.S. receives consulting feeds from BMS, Inc. and is an employee of Insitro, inc. C.M.U. has, as cancer center director, oversight over research funded by several pharmaceutical companies, but has not received funding directly herself.

light-to-moderate drinkers relative to non-drinkers and heavy drinkers has only recently been described and is poorly understood. As a possible explanation, the increased risk of CRC in non-drinkers may be due to residual confounding because some of these individuals may abstain from or stop drinking for reasons related to CRC risk factors or health status, including alcoholism. In fact, the McNabb et al. manuscript describing the J-shape explicitly states that the observed inverse association could be explained by residual confounding or chance [1]. Another possibility is that light-to-moderate drinking has a protective effect on risk of CRC, even though heavier consumption is detrimental. However, this hypothesis is only supported by very preliminary evidence of an anti-inflammatory effect of light-to-moderate drinking on the colon in rats [3,4] and of low levels of ethanol exposure upregulating liver detoxification enzymes [5,6], so future research is needed to explore any possible protective effects.

Given this complex relationship, it is possible that there are single nucleotide polymorphisms (SNPs) that affect only non-drinkers or heavy drinkers or that known loci have unknown interactions with alcohol consumption that would be difficult to detect in genome-wide association studies (GWAS) of CRC. In fact, common SNPs identified through GWAS and hereditary syndromes explain less than half of the roughly 25% of CRCs that aggregate in families [7]. Since alcohol consumption is widespread in the US population and there are known variants in genes like *ADH* and *ALDH* that have strong effects on alcohol metabolism [8], SNPs that have important interactions with alcohol may help fill in this missing heritability [9]. In addition, variant effects in non-coding regions of the genome may play an important role, through interactions with mechanisms like alcohol-induced epigenetic changes in cancer [10]. To search for important relationships with this established risk factor, we conducted genome-wide interaction analyses to test for SNPs that modify the effects of alcohol consumption on risk of CRC, including a novel hybrid two-step approach that aims to improve statistical power.

Materials and Methods

Study population

We pooled individual level genomic and epidemiological data from studies participating in the Colon Cancer Family Registry (CCFR), the Colorectal Transdisciplinary Study (CORECT), and the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO). Study details have been previously published [11-13] and can be found in Supplementary Table 1. For cohort studies, nested case-control sets were assembled using risk-set sampling. Controls were matched on factors such as age, sex, race, and enrollment date or trial group, when applicable. Colorectal adenocarcinoma cases were confirmed by medical records, pathological reports, or death certificate information. For the small subset of advanced adenoma cases, matched controls displayed polyp-free sigmoidoscopy or colonoscopy at the time of adenoma selection. All participants gave written informed consent and studies were approved by their respective Institutional Review Boards.

Analyses were limited to individuals of European ancestry based on self-reported race and clustering of principal components with 1000 Genomes EUR superpopulation, yielding an initial sample size of 96,735. There were approximately 3,660 participants excluded

from our main analysis based on ancestry. We excluded studies based on availability of alcohol consumption information, in addition to studies whose populations lacked sufficient variability in alcohol intake levels, were matched on smoking status, or studies where participants had a history of adenomas at baseline (n=19,259). We further excluded samples that showed cryptic relatedness, were duplicates with lower genotyping quality, had genotyping or imputation errors, or were age outliers (n=3,377), creating a final sample size of 74,099.

Exposure Definition

Demographic and environmental risk factor information was self-reported either at in-person interviews or via structured questionnaires. Harmonization of alcohol intake information consisted of a multi-step procedure performed at Fred Hutchinson Cancer Research Center, which is the GECCO coordinating center [14]. Briefly, common data elements (CDEs) were defined *a priori*. Study questionnaires and data dictionaries were examined and, through an iterative process of communication with data contributors, elements were mapped to these CDEs. Definitions, permissible values, and standardized coding were implemented into a single database via SAS and T-SQL. Resulting data were checked for errors and outlying values within and between studies.

Food frequency questionnaires and diet histories were used to ascertain alcohol intake and other risk factors at the reference time, typically ranging from three months to two years prior to diagnosis for case-control studies and at enrollment for cohort studies (Supplementary Text 1). The harmonized alcohol intake variable is expressed as grams per day, and is categorized into three groups: non-drinkers (≤1 g/day; we did not set this to 0 as some studies included small amounts of alcohol intake from fermented foods), light-to-moderate drinkers (>1 to ≤ 28 g/day), and heavy drinkers (>28 g/day) [15]. To account for the potentially disparate biological mechanisms driving the J-shaped association between alcohol use and CRC, we conducted separate genome-wide interaction scans: non-drinkers vs. light-to-moderate drinkers and heavy drinkers vs. light-to-moderate drinkers. Light-to-moderate drinkers serve as the reference group for both scans as they have the lowest risk of CRC.

Genotyping and Imputation

Details on genotyping and quality control have been previously published [11]; genotyping platforms used are summarized in Supplementary Table 1. Briefly, genotyped SNPs were excluded based on call-rate less than 95-98%, lack of Hardy Weinberg equilibrium with p-value less than $1x10^{-4}$, discrepancies between reported and genotypic sex, and discordant calls between duplicates. Autosomal SNPs of all studies were imputed to the Haplotype Reference Consortium r1.1 (2016) reference panel via the Michigan Imputation Server [16] and converted into a binary format for data management and analyses using R package *BinaryDosage* (Morrison 2019). We filtered imputed SNPs based on a pooled MAF greater than or equal to 1% and imputation accuracy of r^2 greater than 0.80. After imputation and quality control, a total of over 7.2 million common SNPs were used.

Statistical analysis

Interaction Tests—To evaluate main effects, we used logistic regression models adjusted for age at the reference time, sex, and total energy consumption (kcal/day) and stratified by study. Study-specific results were combined using random-effects meta-analysis models using the Hartung-Knapp method to obtain summary odds ratios (ORs) and 95% confidence intervals (CIs) across studies[17]. Random effects were used given the large number of studies and possible heterogeneity of associations[17]. We calculated the heterogeneity p-values using Cochran's Q statistics [18], and funnel plots were used to identify studies with outlying ORs. These analyses were performed using R package *meta* [19] (REF). [19]

We performed genome-wide interaction scans using the R package *GxEScanR*, which implements several interaction testing methods [20], including traditional logistic regression GxE and joint tests of association, as described below. Imputed SNP dosages were modeled as continuous variables [21]. For the purposes of this study, E refers to alcohol exposure, G refers to a SNP included in the genome-wide tests, D refers to CRC disease status, and C refers to a set of adjustment covariates. To test for multiplicative scale interaction, we fit conventional logistic regression models augmented with an interaction term of the form $logit(Pr(D = 1 \text{ V } G)) = \beta_0 + \beta_G G + \beta_E E + \beta_{GxE} GxE + \beta_C C$ and tested $H0{:}\beta_{GxE} = 0$. The quantity $exp(\beta_{GxE}) = OR\{GxE\}$ captures departure from multiplicative associations of E and G on D. The models were adjusted for age at the reference time, sex, study, total energy consumption (kcal/day), and the first three principal components from EIGENSTRAT to account for potential population substructure. For any significant findings, we conducted a sensitivity analysis stratified by sex and tumor site and adjusted for BMI, diabetes, education level, ever smoking, and study and sex-specific quartiles of red meat, fruit, and vegetable consumption. Age at reference time was missing for a single participant and was median imputed by study for cases and controls separately. For the remaining variables, we fit models using only subsets of individuals with available covariate information (complete case analysis). Total energy consumption was imputed for 25,247 individuals using study and sex specific means; missingness was 18% across all studies excluding UK Biobank. For UK Biobank, imputation was not feasible because energy intake information was missing for all individuals in this study. Consequently, in order to retain UK Biobank in the analysis, total energy intake was set to 0.. For the 2df joint test, we used likelihood ratio tests to jointly test $H0{:}\beta_G = \beta_{GxE} = 0$, $df = 2$. To accommodate E|G associations, we also extended this to a 3df likelihood ratio test to jointly test $H0{:}\beta_G = \beta_{GxE} = \delta_G = 0$, $df = 3$, where δ_G represents the association between G and E in a combined case-control sample [22,23]. We report two-sided p-values calculated from these likelihood-ratio tests, and consider a p-value of less than $5x10^{-8}$ significant and of less than $5x10^{-6}$ suggestive.

We also implemented a hybrid two-step method that prioritizes potential interaction loci by weighting GxE tests (step 2) based on the ranks of an independent test statistic, in this case the genetic main effects on CRC (step 1). Our approach modifies the original weighted hypothesis testing framework, which uses step 1 ranks to prioritize and partition SNPs into exponentially larger bins of fixed sizes (based on an initial bin size of 5 and an overall significance level of 0.05) and increasingly more stringent step 2 significance thresholds [24,25]. A limitation of the original approach is that the top bins are often filled with correlated

markers from the same loci in analyses of imputed SNPs. To address this issue, our approach accommodates bins of varying sizes while properly controlling for type I error. Specifically, SNPs are partitioned into bins based on step 1 p-value thresholds *in expectation*, which were calculated using the original predetermined bin sizes and assumed uniform distribution of 1 million independent tests. For step 2 GxE testing, we accounted for the influx of correlated markers into each bin by correcting for the effective number of tests, estimated using principal component analysis performed on bin specific genotype correlation matrices [26]. This modification alleviates multiple testing burden and improves statistical power, while maintaining an overall type I error rate of 0.05. We also estimated stratified ORs by modeling interactions between alcohol intake and posterior genotype probabilities.

Relevant regional plots were generated using the command line version (Standalone) of LocusZoom v1.3 [27]. Measures of linkage disequilibrium (LD) were estimated using study population controls. Possible eQTL relationships were explored using the Genotype-Tissue Expression (GTEx V8) and the University of Barcelona and University of Virginia genotyping and RNA sequencing project (BarcUVa-Seq) [28] datasets. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal on April 14, 2020 and dbGaP accession number phs000424.vN.pN on April 15, 2020. The most promising eQTL-gene association was tested in a subset of 35 human normal colon 3D organoid lines from an ongoing study in which lines were grown and expression was measured as described for the control condition in Devall et al. [29]; lines were genotyped on the OncoArray beadchip, and the variant of interest was imputed with an r^2 of 0.98 using the TOPMed reference panel [30]. We then tested predicted expression of the eQTL-associated gene of interest for an interaction with alcohol consumption in data from the three consortia involved in this study (Supplementary Text 2).

Prediction of regulatory impact of candidate non-coding variants—We used ATAC-seq, DNASE-seq, H3K27ac histone ChIP-seq, and H3K4me1 histone ChIP-seq datasets of primary tissue from healthy colon and tumor primary tissue samples from Scacheri et al. [31], as well as from three CRC cell lines (SW480, HCT116, COLO205). These datasets were processed through ENCODE ATAC-seq/DNASE-seq [32] and histone ChIP-seq pipelines [33] to perform alignment and peak calling. Dataset sources are indicated in Supplementary Table 2. −log10(*p*-value) tracks were extracted from the MACS2 step of the pipeline for visualization in genome browsers. Irreproducible Discovery Rate (IDR) [34] peak calls for ATAC-seq and DNASE-seq datasets, as well as naive overlap peak calls for histone ChIP-seq datasets, were determined from the ENCODE pipelines. The pyGenomeTracks [35] software package was used to visualize chromatin accessibility across the functional datasets and to plot −log10(p-value) signal tracks. Peaks across samples from the same assay were concatenated across datasets, cropped to within 200 bp centered on the peak summit, and merged using bedtools [36] merge.

Gapped *k*-mer support vector machine models (LS-GKM) (v0.1.0) with a center-weighted GKM kernel were trained to classify chromatin accessible regions against genomic background regions as a function of their underlying DNA sequences [37]. Default parameters were utilized. Support vector machines (SVMs) were trained via 10-fold cross-validation, where groups of chromosomes were split into folds (Supplementary Table 3). Separate SVM

models were trained on DNase-seq data from Supplementary Table 2 with samples pooled across assays as described above [31]. For each biosample, the SVMs were trained on 120,000 genomic regions. Positively labeled regions were 1 kb DNA sequences centered on the summits of the 60,000 most significant DNase-seq peaks from MACS2+IDR, and 60,000 1 kb negatively labeled sequences were randomly sampled from the GRCh38 reference genome such that they did not overlap IDR and naive overlap DNase-seq peaks and were matched for GC content to the positive regions.

The resulting trained models for each of the five DNASE-seq datasets were used to score all variants on the Haplotype Reference Consortium (HRC) imputed panel (n=39,117,106). For each SNP along the HRC panel, we centered a 1 kb sequence interval and obtained SVM model predictions for the reference and alternate alleles. The difference in model predictions of accessibility (prediction for alternate allele - prediction for reference allele) are the in-silico mutagenesis scores (ISM), or SNP effect scores. We confirmed that the ISM scores for the HRC panel were normally distributed using the Kolmogorov-Smirnov and Shapiro Wilkes tests (p-values>0.10) and derived Z -scores. Variants with ISM scores greater than 1.65 or less than −1.65, representing a 90% confidence interval, were determined to have significant effects. A single score was obtained for each HRC SNP by taking the maximum of the absolute values of the GKMexplain delta scores across the five models.

The lead GWAS SNP rs11190164 was LD-expanded (500 kb window, r^2 thresholded at 0.20) using PLINK (1.9) [38] based on the 1000 genomes phase 3 fileset from the cog-genomics site (https://www.cog-genomics.org/plink/2.0/resources#1kg_phase3) [39], which was filtered to separate individuals of CEU ancestry. Using the SVM models trained on each of the five DNase-seq datasets, we scored the LD-expanded rs11190164 locus and predicted ISM effects on chromatin accessibility. We further inferred the contribution scores of each nucleotide in the input sequences to the output prediction of the SVM models using the GKMexplain algorithm [40]. For each sequence containing a candidate variant, we computed GkmExplain scores for the sequence containing the reference allele and the sequence containing the alternate allele. For each candidate variant, a deltaGKMexplain score was computed by subtracting the GKMexplain score for the 1 kb vector of GKMexplain scores of the sequence with the reference allele from the 1 kb vector of GKMexplain scores of the sequence with the alternate allele. The TomTom algorithm [41] was used to identify likely motif matches for subsequences with high deltaGKMexplain scores. The support vector machine LS-GKM + GKMexplain workflow source code is available on github: https://github.com/kundajelab/SVM_pipelines.

Candidate functional variants were annotated with the 18-state ChromHMM annotations [42] across 218 cell types from the Roadmap Atlas [43] (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/). Bedtools intersect was utilized to identify overlaps between candidate functional SNPs and regions of enhancer activity in cell types associated with CRC.

Results

As initial steps, we examined the characteristics of participants included in our interaction tests. Cases were older, had higher BMI and energy intake, more frequently had a family history of CRC, had lower levels of education, and were more likely to ever smoke cigarettes (Table 1). We then confirmed the previously reported main effect relationships between alcohol consumption and CRC, where non-drinking (OR=1.13, 95% CI=1.05-1.21) and heavy drinking (OR=1.34, 95% CI=1.23-1.45) were associated with increased risk as compared to light-to-moderate drinking (Supplementary Figure 1). The association between non-drinking and CRC risk was similar across tumor sites, and the heavy drinking association was weakest for proximal colon cancer (OR=1.26) and strongest for distal (OR=1.39) and rectal colon cancer (OR=1.44). We observed substantial heterogeneity in the association between non-drinking and CRC across studies (I^2=66%, p-value<0.001). This observation is consistent with the fact that the reason for abstaining from alcohol and the composition of never, former, and occasional drinkers in the non-drinking group both affect risk of CRC and vary across study populations. This heterogeneity was not observed for the association between heavy drinking and CRC.

Using the traditional genome-wide GxE tests of the interaction, we did not identify a significant interaction between any SNP and alcohol consumption. For the non-drinking as compared to light-to-moderate drinking GxE, there was a suggestive interaction in the 10q24.2/*COX15* region previously associated with CRC (Figure 1A) [12,13]. There were also suggestive interactions from the heavy drinking as compared to light-to-moderate GxE, but none in regions previously identified by GWAS of CRC (Figure 1C). The joint 2-df tests identified SNPs with known CRC associations, and the joint 3-df tests additionally identified SNPs with known alcohol consumption associations; however, no novel GxE interaction was discovered.

We also conducted a hybrid two-step method to test for interactions, which yielded a statistically significant finding in the same 10q24.2/*COX15* locus that had a suggestive GxE interaction. For non-drinkers as compared to light-to-moderate drinkers, there were 13 SNPs with pairwise $r^2 > 0.90$ in the 10q24.2/*COX15* region that showed a statistically significant interaction on risk of CRC (Figure 1B). This procedure was null for heavy as compared to light-to-moderate drinking (Figure 1D). As shown in the regional association plot, the lead SNP with the most significant interaction p-value in the region was rs2300985 (Figure 2). A stratified analysis of the lead SNP illustrates the observed interaction, showing that the A allele of rs2300985 was associated with a higher risk of CRC compared to the G reference allele only in light-to-moderate drinkers; the association was null in non-drinkers and in heavy drinkers (Table 2). We observed a dose-response relationship in light-to-moderate drinkers, where the OR for one copy of the rs2300985 A allele was 1.11 (95% CI=1.06-1.17) and was 1.22 (95% CI=1.14-1.31) for two copies of the A allele (Table 2). Since light-to-moderate drinkers are the reference group, the OR for the interaction term in the pooled GxE was inverse (OR=0.89, 95% CI=0.84-0.94, p-value=1.16 x 10^{-6}). The forest plot illustrates an acceptable level of heterogeneity for the interaction OR across studies and no substantial difference between cohort and case-control studies (Supplementary Figure 2). The interaction term was similar in analyses stratified by sex and

tumor site, though it was weakest in proximal colon cases (OR=0.92) and strongest in distal colon cases (OR=0.87). This result withstood a sensitivity analysis additionally adjusted for BMI, diabetes, education level, ever smoking, as well as study and sex-specific quartiles of red meat, fruit, and vegetable consumption (OR=0.89, 95% CI=0.84-0.94). Additional adjustment for physical activity and post-menopausal hormone replacement therapy use restricted our sample size to 14,948 women and produced an odds ratio of 0.95.

The initial GWAS that discovered the 10q24.2/*COX15* locus identified rs11190164 as the most significantly associated with CRC risk [13]. The lead SNP from our interaction analyses (rs2300985) was highly correlated with rs11190164 at an r^2 of 0.59. We LD-expanded our candidate set of variants to include 158 SNPs, including rs2300985, in a 500 kb window around the rs11190164 lead GWAS SNP based on an r^2 greater than 0.20 in the 1 KG Phase 3 EUR population. We integrated functional chromatin profiling data in healthy colon, CRC tumor tissue, and three cell-lines (SW480, HCT116, COLO205) with machine learning models of regulatory DNA sequence to prioritize putative causal regulatory variants in this locus. For each candidate variant, we used gapped k-mer support vector machine (gkmSVM) models trained on DNase-seq data from the five CRC-relevant biosamples to predict its allelic effect on chromatin accessibility in each biosample (Figure 3). As expected, most of the candidate variants were predicted to have no significant allelic effects on chromatin accessibility. However, the models predicted the rs1318920 variant as a putative causal variant based on a significant difference in predicted chromatin accessibility between the reference C and alternate T allele (ISM score=−1.86, p-value=0.02 in healthy colon; ISM score=−2.22, p-value=0.007 in CRC tumor; ISM score=−1.79, p-value=0.02 in COLO205). The rs1318920 SNP had an association p-value of 6.9×10^{-5} in our GWAS of CRC and an r^2 of 0.60 with both the lead GWAS SNP rs11190164 and the lead interaction SNP rs2300985. The characteristics of the three SNPs of interest in the 10q24.2/COX15 region are described in Supplementary Table 4, which also verifies that their allele frequencies did not differ substantially by category of alcohol intake.

In order to further explore the regulatory sequence features disrupted by each of the candidate variants, we used the GkmExplain method to infer the contribution of each nucleotide in the 1000 bp sequences containing the reference and alternate allele to the predicted chromatin accessibility from the gkmSVM models. GkmExplain analysis of rs1318920, rs11190164 (the lead GWAS SNP), and rs2300985 (the lead interaction SNP) supported the prediction of a strong allelic effect specifically for rs1318920 in healthy tissue, tumor tissue, and the COLO205 cell line (Figure 4A-C). The C allele of the rs1318920 variant was predicted to significantly (p-value=4.2×10^{-5}) amplify the contribution scores of an overlapping subsequence [TTTGGACTTTGACC] relative to the T allele. This subsequence is a strong match to the known binding motif of the Hepatocyte Nuclear Factor 4α (HNF4α) transcription factor. The rs1318920 variant was also found to lie within 50 bp of the overlapping DNase-seq peak summits, which are the locations with maximal signal; this additionally supports its strong effect size via motif disruption (Figure 5). Integrative chromatin state annotations from ChromHMM [42] across 218 biosamples revealed that rs1318920 falls within a putative regulatory element that is in an active enhancer state marked by enhancer-associated H3k27ac and H3K4me1 specifically in colorectal tissues (Supplementary Figure 3). In contrast to rs1318920, the lead GWAS SNP rs11190164 and

the lead interaction SNP rs2300985 did not have any supporting evidence for functional effects.

Two independent sources of expression quantitative trait loci (eQTLs) expand on the regulatory role of rs1318920. rs1318920 is an eQTL in the GTEx v8 compendium that influences the expression of *EBAG9P1*, *ENTPD7*, and *RP11-85A1.3* in brain, cultured fibroblast, esophageal, or nerve tissues. rs1318920 is also a suggestive eQTL in normal colon tissue from the BarcUVa-Seq study [28] that regulates expression of *ENTPD7*, where the T alternate allele is associated with increased expression (β=0.11, p-value=3.5 x 10^{-3}). Based on the BarcUVa-Seq result, we checked for and similarly observed a positive association between the T allele of rs1318920 and *ENTPD7* expression in human normal colon 3D organoids (β=0.59, p-value=0.004). Exploring these results further, we detected a statistically significant interaction between standardized predicted expression of *ENTPD7* and non-drinking (p-value=0.007) in our data from the involved consortia; the interaction was positive, but non-significant for heavy drinking (p-value=0.42) (Supplementary Table 5).

Discussion

We conducted a genome-wide interaction study (GWIS) of CRC and discovered a possible interaction between alcohol consumption and genetic variants in the 10q24.2/*COX15* region. Specifically, for the lead interaction SNP, we found that the A allele of rs2300985 was associated with an increase in risk of CRC in light-to-moderate drinkers, but was not associated with CRC risk in non-drinkers.

If non-drinking partially captures other risk factors or health status, then our result suggests that those characteristics overwhelm or counteract the effects associated with rs2300985 in non-drinkers. If light-to-moderate drinking is in fact protective, a possible mechanism might be that low levels of ethanol exposure are anti-inflammatory [3,4] or upregulate liver detoxification enzymes that then mitigate other risk factors, while the adverse consequences of alcohol predominate over any hypothesized benefits at high levels of ethanol exposure [5,6]. Based on our results, the effects associated with rs2300985 may be related to carcinogenesis only when combined with other changes due to light-to-moderate drinking.

Our integrative analysis also suggests rs1318920 as a potentially causal variant in the 10q24.2/*COX15* region that is in LD with both the lead GWAS SNP rs11190164 and the lead interaction SNP rs2300985. rs1318920 is predicted to have a significant allelic effect on chromatin accessibility of a colorectal tissue-specific active enhancer by restoring the combined binding motif of HNF4 α and γ and may have a regulatory effect on expression of *ENTPD7* in colon tissue. As a potential connection to alcohol consumption, there is evidence from mouse and cell-line experiments that HNF4α DNA binding inhibits alcoholic steatosis [44] and may prevent alcoholic liver disease, which is a possible risk factor for CRC [45]. As a result, we hypothesize that rs1318920 is the causal variant driving the observed increased risk of CRC and that its effects on CRC may be affected by alcohol consumption. This finding warrants future work to confirm the functional relevance of rs1318920 in

cancer cell lines, including results from a luciferase assay demonstrating the allele-specific enhancer activity that was predicted in the COLO205 cell line.

If confirmed, the possible causal variant suggests a plausible biological mechanism, where the T alternate allele completes the motif that binds to the HNF4α transcription factor. *HNF4A* regulates genes involved in glucose, cholesterol, fatty acid, and amino acid metabolism [46]; it has also been linked to CRC and identified as a potential drug target by Sladek et al. [47]. In this case, the HNF4α binding site is within *COX15* and close to *ENTPD7*. While speculative, the T allele of rs1318920 in the 10q24.2/*COX15* region may restore HNF4α binding and promote CRC, possibly through increased expression of *ENTPD7* in the colon. Interestingly, SNPs located in the 10q24.2/*COX15* region are also *cis*-eQTLs for *EBAG9P1* in naive CD4+ T cells, CD8+ T cells, and T_{REG} immune cells [48], suggesting a potentail impact on CRC risk via an immunomodulating effect. Follow-up analyses are warranted to assess the functional support for our hypotheses and to explore additional plausible mechanisms, including possible pathways through folate deficiency.

In a prior GWIS of alcohol in this consortium [49], we highlighted an interaction between SNPs in the 9q22.32/*HIATL1* region and light-to-moderate drinking as compared to non-drinking. The tag SNP rs9409565 met the p-value threshold of 0.05 for validation, but was not statistically significant in our larger study after adjustment for multiple testing (OR=0.94, p-value=0.01).

Though GxE interactions are difficult to detect, our GWIS benefits from a substantial sample size of 31,874 cases and 42,225 controls, which was only possible through the inclusion of numerous epidemiological studies with detailed risk assessment and dedicated data harmonization efforts over many years. This study is currently among the largest of its kind and, combined with cutting edge statistical methods, allowed us to detect a possible interaction between a known GWAS hit and alcohol consumption. Our detailed evaluation of the J-shaped relationship between alcohol consumption and risk of CRC [1] also ensured that we appropriately modeled alcohol consumption during interaction testing.

Our study also has several limitations. Our cohort, though large, consisted of consortia involving individuals solely from EUR backgrounds, which limits the generalizability of the findings to individuals of non-European ancestry. Our categories of alcohol consumption were not sex-specific; however, we observed similar main effect associations between alcohol consumption and CRC in males and females, and we adjusted our interaction tests for sex. Alcohol consumption was based on intake at the reference time, so our non-drinking category includes former drinkers, and this approach may contribute to residual confounding in the non-drinking group. Given the complexity of the harmonization process and the inconsistent information about past drinking across the large number of diverse studies involved, a sensitivity analysis excluding former drinkers is not feasible. However, to explain our main finding, the presence of former heavy drinkers in the non-drinking group would need to attenuate the association between the rs2300985 A allele and CRC risk more than observed in the heavy drinking group itself and would also need to outweigh the bias away from the null introduced by the presence of former light-to-moderate drinkers. The interaction also survived a sensitivity analysis adjusted for a comprehensive set of

potential confounders, though we were limited to covariates that were harmonized across the consortia and do not have access to all measured variables in each participating study. There were more non-drinkers than heavy drinkers, so we were not as well powered to detect interactions with heavy alcohol consumption.

For the reported interaction, the result from the traditional GxE was only suggestive, especially considering we conducted two GWIS. The modified two-step method that reported 10q24.2/COX15 as statistically significant is a newer approach that controls for multiple testing by estimating the bin-specific effective number of tests. While this approach is computationally more expensive, it addresses an important limitation of prior two-step methods, which do not currently account for correlated markers. Multiple methods exist to calculate the effective number of tests, and the Gao et al. method used in our approach is a comparatively stringent option [26,50]. Finally, we were unable to establish causality with the data available, so the proposed relationships need to be validated experimentally using methods like Perturb-seq coupled with differing alcohol treatment conditions [51,52].

In summary, our results suggest that the association at the known 10q24.2/*COX15* CRC locus is driven by light-to-moderate drinkers. Further, we have identified a putative causal variant in the region with strong evidence for a functional effect, which provides interesting directions for future research involving the link between rs1318920 and CRC and the possible role of alcohol consumption in this mechanism. Though we hope these findings inform future research involving the 10q24.2/*COX15* region and CRC, they should not be used to guide public health recommendations without further validation, functional work, and research in the context of other CRC-associated variants and risk factors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Kristina M Jordahl[1,2], Anna Shcherbina[3,4], Andre E Kim[5], Yu-Ru Su[2], Yi Lin[2], Jun Wang[5], Conghui Qu[2], Demetrius Albanes[6], Volker Arndt[7], James Baurley[8,9], Sonja I Berndt[6], Stephanie A Bien[2], D Timothy Bishop[10], Emmanouil Bouras[11], Hermann Brenner[7,12,13], Daniel D Buchanan[14,15,16], Arif Budiarto[8,17], Peter T Campbell[18], Robert Carreras-Torres[19], Graham Casey[20], Tjeng Wawan Cenggoro[8], Andrew T Chan[21,22,23,24,25,26], David V Conti[27], Christopher H Dampier[20], Matthew Devall[20], Virginia Díez-Obrero[19,28,29,30], Niki Dimou[31], David A. Drew[32], Jane C Figueiredo[33,34], Steven Gallinger[35], Graham G Giles[36,37,38], Stephen B Gruber[27], Andrea Gsur[39], Marc J Gunter[40], Heather Hampel[41], Sophia Harlid[42], Tabitha A Harrison[2], Akihisa Hidaka[2], Michael Hoffmeister[7], Jeroen R Huyghe[2], Mark A Jenkins[37], Amit D Joshi[23,25], Temitope O Keku[43], Susanna C Larsson[44], Loic Le Marchand[45], Juan Pablo Lewinger[34], Li Li[46], Bharuno Mahesworo[8], Victor Moreno[28,30,47,48], John Morrison[34], Neil Murphy[40], Hongmei Nan[49,50], Rami Nassir[51], Polly A Newcomb[1,2], Mireia Obon-Santacana[52], Shuji Ogino[25,53,54,55], Jennifer Ose[56,57], Rish K Pai[58], Julie R Palmer[59], Nikos Papadimitriou[31], Bens

Pardamean[8], Anita R Peoples[60,61], Paul D P Pharoah[62], Elizabeth A Platz[63], John D Potter[2], Ross L Prentice[2], Gad Rennert[64,65,66], Edward Ruiz-Narvaez[67], Lori C Sakoda[2,68], Peter C Scacheri[69], Stephanie L Schmit[70,71], Robert E Schoen[72], Martha L Slattery[73], Mariana C. Stern[27], Catherine M Tangen[74], Stephen N Thibodeau[75], Duncan C Thomas[27], Yu Tian[76,77], Konstantinos K Tsilidis[11,78], Cornelia M Ulrich[60,61], Franzel JB van Duijnhoven[79], Bethany Van Guelpen[42,80], Kala Visvanathan[63], Pavel Vodicka[81,82,83], Emily White[1,2], Alicja Wolk[44], Michael O Woods[84], Anna H Wu[34], Natalia Zemlianskaia[27], Jenny Chang-Claude[76,85], W James Gauderman[27], Li Hsu[2,86], Anshul Kundaje[3,4], Ulrike Peters[1,2]

Affiliations

[1]Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, USA

[2]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

[3]Department of Genetics, Stanford University, Stanford, California, USA

[4]Department of Computer Science, Stanford University, Stanford, California, USA

[5]Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

[6]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA

[7]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany

[8]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia

[9]BioRealm LLC, Walnut, California, USA

[10]Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK

[11]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

[12]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany

[13]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

[14]Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of Melbourne, Parkville, Victoria 3010 Australia

[15]University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Parkville, Victoria 3010 Australia

[16]Genetic Medicine and Family Cancer Clinic, The Royal Melbourne Hospital, Parkville, Victoria, Australia

[17]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

[18]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, Georgia, USA

[19]Colorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain

[20]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA

[21]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

[22]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

[23]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

[24]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

[25]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA

[26]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA

[27]Department of Preventive Medicine & USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

[28]Oncology Data Analytics Program, Catalan Institute of Oncology-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain

[29]Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain

[30]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain

[31]Section of Nutrition and Metabolism, International Agency for Research on Cancer, Lyon, France

[32]Clinical & Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[33]Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[34]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

[35]Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada

[36]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia

[37]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia

[38]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia

[39]Institute of Cancer Research, Department of Medicine I, Medical University Vienna, Vienna, Austria

[40]Nutrition and Metabolism Section, International Agency for Research on Cancer, World Health Organization, Lyon, France

[41]Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, USA

[42]Department of Radiation Sciences, Oncology Unit, Umeå University, Umeå, Sweden

[43]Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, North Carolina, USA

[44]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

[45]University of Hawaii Cancer Center, Honolulu, Hawaii, USA

[46]Department of Family Medicine, University of Virginia, Charlottesville, Virginia, USA

[47]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

[48]ONCOBEL Program, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain

[49]Department of Epidemiology, Richard M. Fairbanks School of Public Health, Indianapolis, Indiana, USA

[50]IU Melvin and Bren Simon Cancer Center, Indiana University, Indianapolis, Indiana, USA

[51]Department of Pathology, School of Medicine, Umm Al-Qura’a University, Saudi Arabia

[52]Unit of Nutrition, Environment and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO-IDIBELL), Avda Gran Via Barcelona 199-203, 08908L'Hospitalet de Llobregat, Barcelona, Spain

[53]Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA

[54]Harvard Medical School, Boston, Massachusetts, USA

[55]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[56]Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah

[57]Department of Population Health Sciences, University of Utah, Salt Lake City, Utah

[58]Department of Laboratory Medicine and Pathology, Mayo Clinic Arizona, Scottsdale, Arizona, USA

[59]Slone Epidemiology Center at Boston University, Boston, MA, USA

[60]Huntsman Cancer Institute, Salt Lake City, Utah, USA

[61]Department of Population Health Sciences, University of Utah, Salt Lake City, Utah, USA

[62]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[63]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[64]Department of Community Medicine and Epidemiology, Lady Davis Carmel Medical Center, Haifa, Israel

[65]Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel

[66]Clalit National Cancer Control Center, Haifa, Israel

[67]Department of Nutritional Sciences, University of Michigan School of Public Health, Ann Arbor, Michigan, USA

[68]Division of Research, Kaiser Permanente Northern California, Oakland, California, USA

[69]Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, Ohio, USA

[70]Genomic Medicine Institute, Cleveland Clinic, Cleveland, Ohio, USA

[71]Population and Cancer Prevention Program, Case Comprehensive Cancer Center, Cleveland, OH

[72]Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA

[73]Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA

[74]SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

[75]Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA

[76]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

[77]School of Public Health, Capital Medical University, Beijing, China

[78]Department of Epidemiology and Biostatistics, Imperial College London, School of Public Health, London, UK

[79]Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, The Netherlands

[80]Wallenberg Centre for Molecular Medicine, Umeå University, Umeå, Sweden

[81]Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic

[82]Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University, Prague, Czech Republic

[83]Faculty of Medicine and Biomedical Center in Pilsen, Charles University, Pilsen, Czech Republic

[84]Memorial University of Newfoundland, Discipline of Genetics, St. John's, Canada

[85]University Medical Centre Hamburg-Eppendorf, University Cancer Centre Hamburg (UCCH), Hamburg, Germany

[86]Department of Biostatistics, University of Washington, Seattle, Washington, USA

Acknowledgements

Genotyping services were provided by the Center for Inherited Disease Research (CIDR).

Cancer data was provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health. This research has been conducted using the UK Biobank Resource under Application Number 8614. We would also like to acknowledge data from VITAL, WHI.

We thank all participants and cooperating clinicians, and everyone who provided excellent technical assistance from the following organizations, registries, and consortia: Colon CFR; Seattle Colon Cancer Family Registry; Hormones and Colon Cancer study (CORE Studies); CLUE II; CPS-II; Czech Republic CCS; DACHS; EPIC; Harvard cohorts (HPFS, NHS); Channing Division of Network Medicine; Department of Medicine, Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health; Kentucky Cancer Registry; LCCS: We acknowledge the contributions of Jennifer Barrett, Robin Waxman, Gillian Smith and Emma Northwood in conducting this study; PLCO Cancer Screening Trial; District of Columbia Cancer Registry, Georgia Cancer Registry, Hawaii Cancer Registry, Minnesota Cancer Surveillance System, Missouri Cancer Registry, Nevada Central Cancer Registry, Pennsylvania Cancer Registry, Texas Cancer Registry, Virginia Cancer Registry, and Wisconsin Cancer Reporting System. All are supported in part by funds from the Center for Disease Control and Prevention, National Program for Central Registries, local states or by the National Cancer Institute, Surveillance, Epidemiology, and End Results program; SELECT; WHI.

The results reported here and the conclusions derived are the sole responsibility of the authors.

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Funding

Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA137088, R01 CA059045, U01 CA164930, R01201407).

References

1. McNabb S et al. Meta-analysis of 16 studies of the association of alcohol with colorectal cancer. Int. J. Cancer 146, 861–873 (2020). [PubMed: 31037736]
2. Rossi M, Anwar MJ, Usman A, Keshavarzian A & Bishehsari F Colorectal Cancer and Alcohol Consumption—Populations to Molecules. Cancers vol. 10 38 (2018).
3. Pai JK et al. Moderate alcohol consumption and lower levels of inflammatory markers in US men and women. Atherosclerosis 186, 113–120 (2006). [PubMed: 16055129]
4. Klarich DS et al. Effects of moderate alcohol consumption on gene expression related to colonic inflammation and antioxidant enzymes in rats. Alcohol 61, 25–31 (2017). [PubMed: 28599714]
5. Gunji T et al. Modest alcohol consumption has an inverse association with liver fat content. Hepatogastroenterology 59, 2552–2556 (2012). [PubMed: 22534544]
6. Alatalo PI et al. Effect of moderate alcohol consumption on liver enzymes increases with increasing body mass index. Am. J. Clin. Nutr 88, 1097–1103 (2008). [PubMed: 18842799]
7. Schubert SA, Morreau H, de Miranda NFCC & van Wezel T The missing heritability of familial colorectal cancer. Mutagenesis 35, 221–231 (2020). [PubMed: 31605533]
8. Morozova TV, Mackay TFC & Anholt RRH Genetics and genomics of alcohol sensitivity. Molecular Genetics and Genomics vol. 289 253–269 (2014). [PubMed: 24395673]
9. van Ijzendoorn MH et al. Gene-by-environment experiments: a new approach to finding the missing heritability. Nature reviews. Genetics vol. 12 881; author reply 881 (2011).
10. Dumitrescu RG Alcohol-Induced Epigenetic Changes in Cancer. Methods Mol. Biol 1856, 157–172 (2018). [PubMed: 30178251]
11. Huyghe JR et al. Discovery of common and rare genetic risk variants for colorectal cancer. Nat. Genet 51, 76–87 (2019). [PubMed: 30510241]
12. Schmit SL et al. Novel Common Genetic Susceptibility Loci for Colorectal Cancer. JNCI: Journal of the National Cancer Institute 111, 146–157 (2018).
13. Schumacher FR et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat. Commun 6, 7138 (2015). [PubMed: 26151821]
14. Hutter CM et al. Characterization of gene-environment interactions for colorectal cancer susceptibility loci. Cancer Res. 72, 2036–2044 (2012). [PubMed: 22367214]
15. Beuth J & Moss RW Complementary Oncology: Adjunctive Methods in the Treatment of Cancer. (Thieme, 2011).
16. Das S et al. Next-generation genotype imputation service and methods. Nat. Genet 48, 1284–1287 (2016). [PubMed: 27571263]
17. Hartung J & Knapp G A refined method for the meta-analysis of controlled clinical trials with binary outcome. Stat. Med 20, 3875–3889 (2001). [PubMed: 11782040]
18. Cochran WG The Combination of Estimates from Different Experiments. Biometrics vol. 10 101 (1954).
19. Schwarzer G, Carpenter JR & Rücker G Meta-Analysis with R. (Springer, 2015).
20. Morrison J GxEScanR: Run GWAS/GWEIS Scans Using Binary Dosage Files. (2020).
21. Zheng J, Li Y, Abecasis GR & Scheet P A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genet. Epidemiol 35, 102–110 (2011). [PubMed: 21254217]
22. Murcray CE, Lewinger JP & Gauderman WJ Gene-Environment Interaction in Genome-Wide Association Studies. Am. J. Epidemiol 169, 219–226 (2009). [PubMed: 19022827]
23. Gauderman WJ et al. A Unified Model for the Analysis of Gene-Environment Interaction. Am. J. Epidemiol 188, 760–767 (2019). [PubMed: 30649161]
24. Kooperberg C & LeBlanc M Increasing the power of identifying gene × gene interactions in genome-wide association studies. Genet. Epidemiol 32, 255–263 (2008). [PubMed: 18200600]

25. Ionita-Laza I, McQueen MB, Laird NM & Lange C Genomewide Weighted Hypothesis Testing in Family-Based Association Studies, with an Application to a 100K Scan. Am. J. Hum. Genet 81, 607–614 (2007). [PubMed: 17701906]
26. Gao X, Starmer J & Martin ER A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet. Epidemiol 32, 361–369 (2008). [PubMed: 18271029]
27. Pruim RJ et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336–2337 (2010). [PubMed: 20634204]
28. Díez-Obrero V et al. Genetic Effects on Transcriptome Profiles in Colon Epithelium Provide Functional Insights for Genetic Risk Loci. Cell Mol Gastroenterol Hepatol 12, 181–197 (2021). [PubMed: 33601062]
29. Devall M et al. Ethanol exposure drives colon location specific cell composition changes in a normal colon crypt 3D organoid model. Scientific Reports vol. 11 (2021).
30. Taliun D et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299 (2021). [PubMed: 33568819]
31. Cohen AJ et al. Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. Nat. Commun 8, 1–13 (2017). [PubMed: 28232747]
32. Lee J et al. ENCODE-DCC/atac-seq-pipeline: v1.9.1 (2020) doi:10.5281/zenodo.4204092.
33. Lee J, Seth Strattan J, annashcherbina, Kagda M & Maurizio PL ENCODE-DCC/chip-seq-pipeline2: v1.6.1 (2020) doi:10.5281/zenodo.4204129.
34. Li Q, Brown JB, Huang H & Bickel PJ Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics vol. 5 (2011).
35. Lopez-Delisle L et al. pyGenomeTracks: reproducible plots for multivariate genomic data sets. Bioinformatics (2020) doi:10.1093/bioinformatics/btaa692.
36. Quinlan AR BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current Protocols in Bioinformatics vol. 47 (2014).
37. Lee D LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics 32, 2196–2198 (2016). [PubMed: 27153584]
38. Chang CC et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience vol. 4 (2015).
39. Katsnelson A 1000 Genomes Project reveals human variation. Nature (2010) doi:10.1038/news.2010.567.
40. Shrikumar A, Prakash E & Kundaje A Gkmexplain: Fast and Accurate Interpretation of Nonlinear Gapped k-mer SVMs Using Integrated Gradients. doi:10.1101/457606.
41. Gupta S, Stamatoyannopoulos JA, Bailey TL & Noble W Quantifying similarity between motifs. Genome Biology vol. 8 R24 (2007). [PubMed: 17324271]
42. Ernst J & Kellis M Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. 12, 2478–2492 (2017). [PubMed: 29120462]
43. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]
44. Kang X et al. Zinc supplementation reverses alcohol-induced steatosis in mice through reactivating hepatocyte nuclear factor-4alpha and peroxisome proliferator-activated receptor-alpha. Hepatology 50, 1241–1250 (2009). [PubMed: 19637192]
45. Komaki Y, Komaki F, Micic D, Ido A & Sakuraba A Risk of colorectal cancer in chronic liver diseases: a systematic review and meta-analysis. Gastrointest. Endosc 86, 93–104.e5 (2017). [PubMed: 28011280]
46. Stoffel M & Duncan SA The maturity-onset diabetes of the young (MODY1) transcription factor HNF4alpha regulates expression of genes required for glucose transport and metabolism. Proc. Natl. Acad. Sci. U. S. A 94, 13209–13214 (1997). [PubMed: 9371825]
47. Chellappa K, Robertson GR & Sladek FM HNF4α: a new biomarker in colon cancer? Biomark. Med 6, 297 (2012). [PubMed: 22731903]
48. Schmiedel BJ et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell vol. 175 1701–1715.e16 (2018). [PubMed: 30449622]

49. Gong J et al. Genome-Wide Interaction Analyses between Genetic Variants and Alcohol Consumption and Smoking for Risk of Colorectal Cancer. PLoS Genet. 12, e1006296 (2016). [PubMed: 27723779]

50. Li M-X, Yeung JMY, Cherny SS & Sham PC Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum. Genet 131, 747–756 (2012). [PubMed: 22143225]

51. Dixit A et al. Perturb-seq: Dissecting molecular circuits with scalable single cell RNA profiling of pooled genetic screens. Cell 167, 1853 (2016). [PubMed: 27984732]

52. Schraivogel D et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat. Methods 17, 629–635 (2020). [PubMed: 32483332]

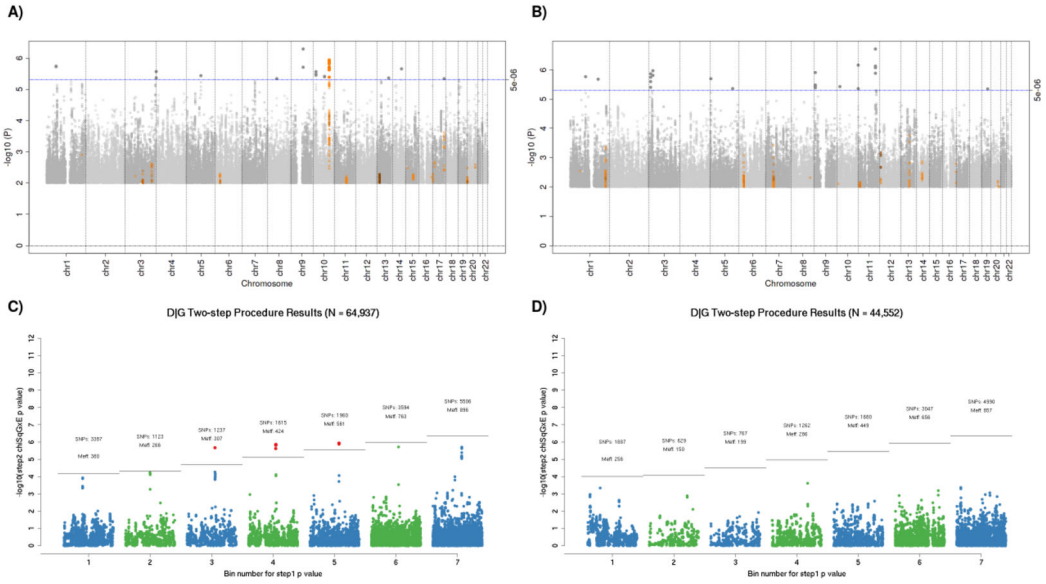

Figure 1.

All analyses are adjusting for age, sex, study site, total energy consumption, and the first three principal components. **A) & C)** Manhattan plots of interaction between genome-wide genetic variants and non-drinking (A) or heavy drinking (C) as compared to light-to-moderate drinking. The blue horizontal line indicates the threshold for suggestive hits (p-value < 5e-6), and SNPs plotted in orange have previously reported associations with colorectal cancer. **B) & D)** Plots of expectation-based partitions adjusted by the number of effective tests in each bin. The gray line indicated the threshold for significance based on the bin specific alpha-threshold (Meff). (B) shows 13 significant SNPs, which are all located in the 10q24.2/*COX15* region. Point colors alternate blue and green for visibility; red points denote statistically significant findings. Abbreviations: SNPs = number of markers included in each bin. Meff = the number of effective tests in each bin after accounting for correlation between SNPs.

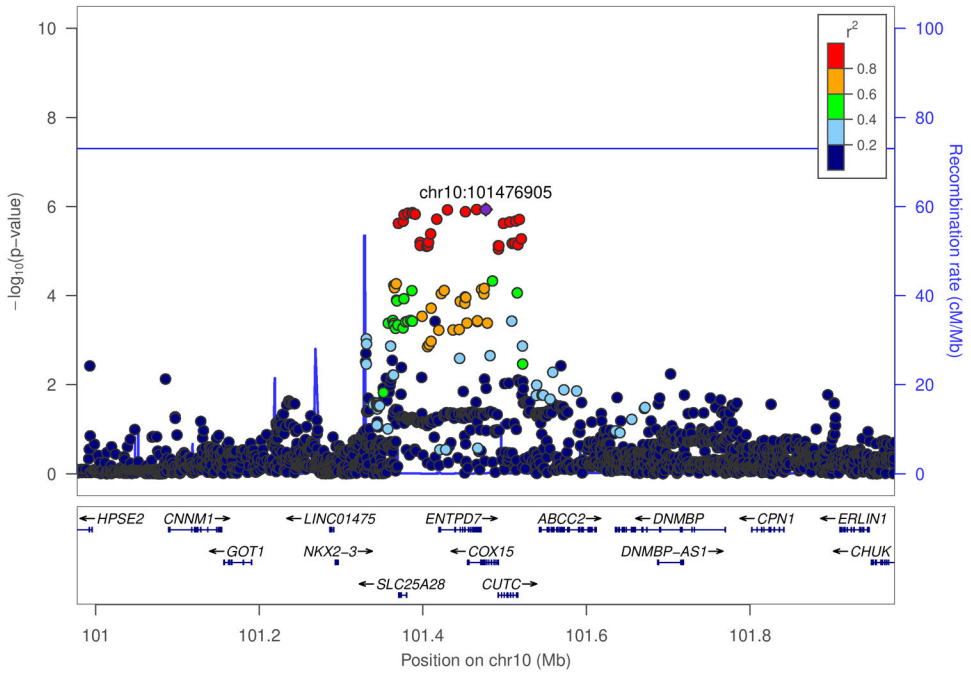

Figure 2.

Regional association plot of SNP and non-drinking vs light-to-moderate drinking interaction $-\log_{10}$ p-values. Result from hybrid two-step analysis of colorectal cancer risk at 10q24.2/*COX15*. rs2300985 is the index SNP as indicated by the purple diamond (GRCh37 coordinates).

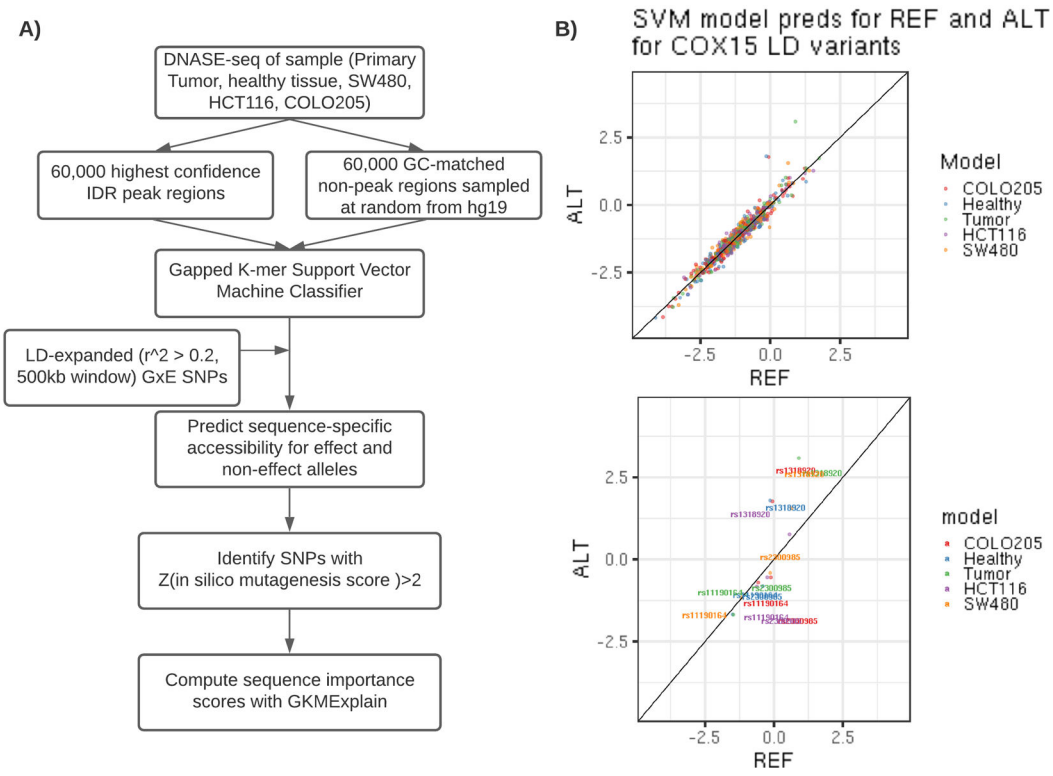

Figure 3.

Support vector machine learning model pipeline to predict functional effects of linked SNPs within the 10q24.2/*COX15* region. **A)** Analysis pipeline for SVM classifier development and linked SNP scoring. **B)** SVM test set predictions for reference and alternate alleles for 158 variants with $r^2 > 0.2$ within 500kb of the *COX15* tagged SNP rs11190164. Bottom panel highlights reference and alternate predictions for rs11191064, rs1318920, and rs2300985.

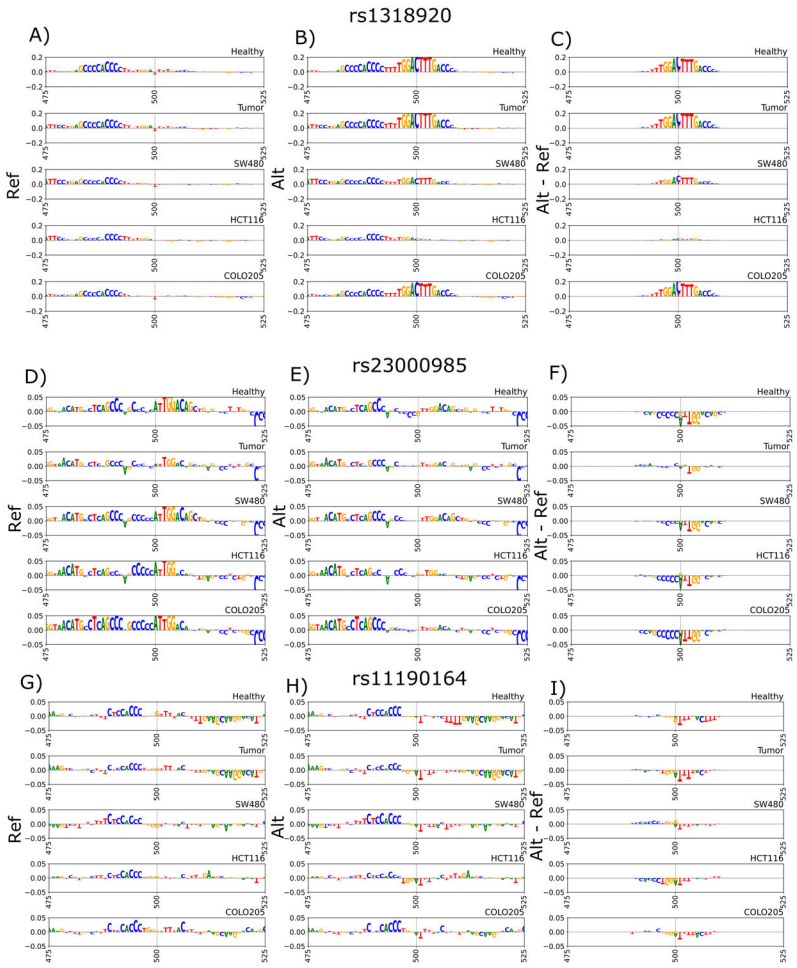

Figure 4.

GkmExplain sequence importance scores within +/− 50 bp of the variants of interest in the 10q24.2/*COX15* region. Scores are derived from SVM models in healthy and tumor primary tissue samples as well as SVM models in cell lines SW480, HCT116, COLO205. **A)** rs1318920 reference allele scores. **B)** rs1318290 alternate allele scores. **C)** rs1318920 alternate allele scores - reference allele scores. **D)** rs2300985 reference allele scores. **E)** rs2300985 alternate allele scores. **F)** rs2300985 alternate allele scores minus reference allele scores. **G)** Tag SNP rs11190164 reference allele scores. **H)** Tag SNP rs11190164 alternate allele scores. **I)** Tag SNP rs11190164 alternate allele scores minus reference allele scores.

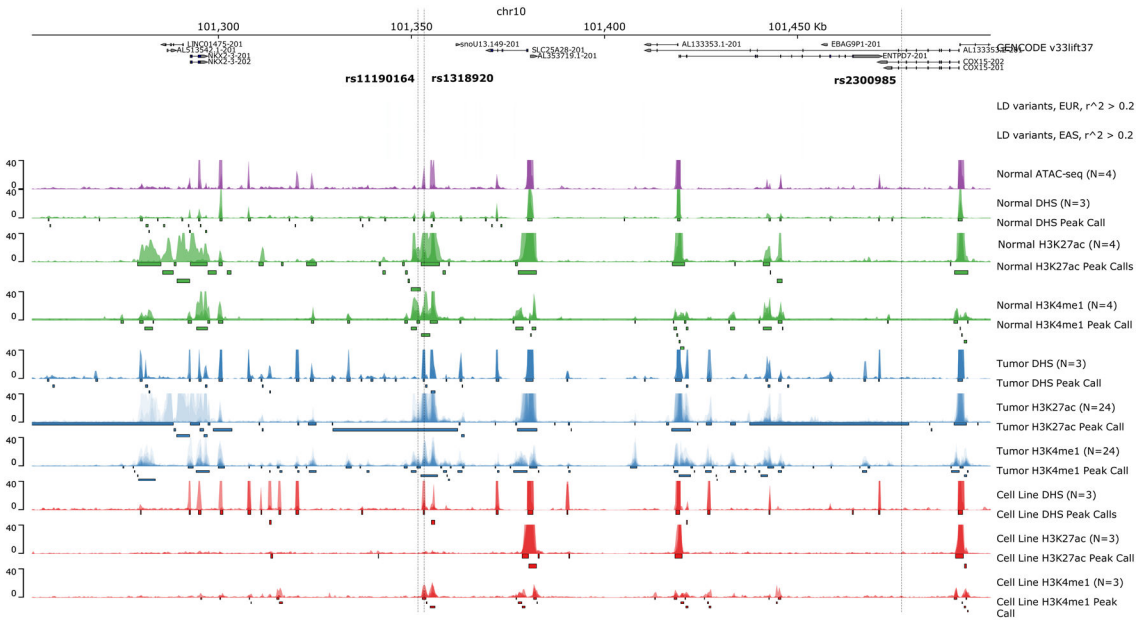

Figure 5.

Chromatin accessibility assays highlighting peaks within the 10q24.2/*COX15* region. Top panel indicates GENCODE reference genes (GRCh37). Variants with r^2 > 0.2 within 500kb of tag SNP rs11190164 are color-coded by r^2 value. LD was calculated for the EUR and EAS populations within phase 3 of the 1000 Genomes (panel 2 and 3 from the top). Healthy ATAC-seq, DNASE-seq, H3K27ac histone ChIP-seq, H3K4me1 histone ChIP-seq p-value bigwigs are indicated in green. The same set of assays for tumor samples are indicated in blue. The same set of assays for cell lines SW480, HCT116, COLO205 are overlaid and indicated in red.

Table 1.

Characteristics of all study participants by case-control status.

	Cases (N=31874)	Controls (N=42225)	P-value
Alcohol consumption [a]			
Light-to-moderate drinkers (>1-28 g/d)	13979 (44 %)	21658 (51 %)	<0.001
Non-drinkers (≤ 1 g/day)	13754 (43 %)	15546 (37 %)	
Heavy drinkers (>28 g/d)	4141 (13 %)	5021 (12 %)	
Age (median imputed)			
Mean (SD)	64.0 (± 10.4)	63.1 (± 9.44)	<0.001
Sex			
Female	15531 (49 %)	21046 (50 %)	0.00269
Male	16343 (51 %)	21179 (50 %)	
Total energy intake (mean imputed) [d]			
Mean (SD)	1910 (± 708)	1970 (± 736)	<0.001
Family history of colorectal cancer			
No	22482 (71 %)	27925 (66 %)	<0.001
Yes	4371 (14 %)	4481 (11 %)	
Missing	5021 (15.8%)	9819 (23.3%)	
BMI			
Mean (SD)	27.4 (± 4.89)	27.0 (± 4.62)	<0.001
Missing	697 (2.2%)	604 (1.4%)	
Education level (highest completed)			
Less than High School	7759 (24 %)	8313 (20 %)	<0.001
High School/GED	6391 (20 %)	6420 (15 %)	
Some College	7651 (24 %)	10780 (26 %)	
College/Graduate School	9011 (28 %)	13587 (32 %)	
Missing	1062 (3.3%)	3125 (7.4%)	
Ever smoking			
No	14284 (45 %)	20496 (49 %)	<0.001
Yes	17093 (54 %)	21089 (50 %)	
Missing	497 (1.6%)	640 (1.5%)	
Type 2 diabetes (ever diagnosed)			
No	26725 (84 %)	37268 (88 %)	<0.001
Yes	3837 (12 %)	3627 (9 %)	
Missing	1312 (4.1%)	1330 (3.1%)	
Total dietary red meat intake [b]			
Q1	7108 (22 %)	10764 (25 %)	<0.001
Q2	8320 (26 %)	11986 (28 %)	
Q3	8088 (25 %)	10910 (26 %)	
Q4	7398 (23 %)	7717 (18 %)	
Missing	960 (3.0%)	848 (2.0%)	

	Cases (N=31874)	Controls (N=42225)	P-value
Total dietary fruit intake [b]			
Q1	8406 (26 %)	10215 (24 %)	<0.001
Q2	9749 (31 %)	11841 (28 %)	
Q3	6832 (21 %)	9923 (24 %)	
Q4	5868 (18 %)	9261 (22 %)	
Missing	1019 (3.2%)	985 (2.3%)	
Total dietary vegetable intake [b]			
Q1	7124 (22 %)	9896 (23 %)	<0.001
Q2	10091 (32 %)	11515 (27 %)	
Q3	7459 (23 %)	10561 (25 %)	
Q4	6248 (20 %)	9326 (22 %)	
Missing	952 (3.0%)	927 (2.2%)	
Physical activity (MET-hr/week) [c]			
Mean (SD)	44.8 (± 64.9)	48.0 (± 70.6)	<0.001
Missing	14547 (45.6%)	16449 (39.0%)	
Post-menopausal hormone replacement therapy use			
No	7510 (24 %)	10605 (25 %)	<0.001
Yes	3827 (12 %)	6032 (14 %)	
Missing	20537 (64.4%)	25588 (60.6%)	
Tumor site			
Distal	8445 (26 %)	0 (0 %)	NA
Proximal	10035 (31 %)	0 (0 %)	
Rectal	8167 (26 %)	0 (0 %)	
Missing	5227 (16.4%)	42225 (100%)	

[a] Non-drinking is treated as missing for the heavy vs. light-to-moderate comparison, and heavy drinking is treated as missing for the non-drinking vs. light-to-moderate comparison. MECC_1 is also excluded from the heavy vs. light-to-moderate comparison, so the heavy drinking interaction analyses involved 247 fewer light-to-moderate drinkers than shown in the table.

[b] Study- and sex- specific quartiles of serving size.

[c] MET defined as 1 kcal/hr/hour. Calculated as the mean +/− 3*(study- and sex- specific mean absolute deviation).

[d] Calculations exclude individuals with missing total energy intake information.

Table 2.

Colorectal cancer associations stratified by genotypes of rs2300985 in the 10q24.2/*COX15* region and by alcohol consumption. A) Stratified associations for rs2300985 genotypes with colorectal cancer within alcohol consumption categories. B) Joint associations for rs2300985 genotypes and alcohol consumption with colorectal cancer across when comparing to light-to-moderate drinkers with the GG genotype.

A)

Genotype at rs2300985	Non-drinkers[a]			Light-to-moderate drinkers[b]			Heavy drinkers[c]		
	No. of Cases[e]	No. of Controls[e]	Odds Ratio (95% CI)[d]	No. of Cases[f]	No. of Controls[f]	Odds Ratio (95% CI)[e]	No. of Cases[g]	No. of Controls[g]	Odds Ratio (95% CI)[e]
GG	5,366	5,747	1 (ref)	4,806	7,804	1 (ref)	1,349	1,639	1 (ref)
GA	6,324	7,369	0.96(0.91-1.01)	6,678	10,266	1.11(1.06-1.17)	2,057	2,496	1.05(0.95-1.16)
AA	2,064	2,430	0.98 (0.91-1.06)	2,495	3,588	1.22(1.14-1.31)	735	886	1.06(0.93-1.21)

B)

Genotype at rs2300985	Non-drinkers[a]			Light-to-moderate drinkers[b]			Heavy drinkers[c]		
	No. of Cases[e]	No. of Controls[e]	Odds Ratio (95% CI)[d]	No. of Cases[f]	No. of Controls[f]	Odds Ratio (95% CI)[e]	No. of Cases[g]	No. of Controls[g]	Odds Ratio (95% CI)[e]
GG	5,366	5,747	1.28 (1.21-1.35)	4,806	7,804	1 (ref)	1,349	1,639	1.45 (1.33-1.58)
GA	6,324	7,369	1.23 (1.17-1.30)	6,678	10,266	1.11 (1.06-1.17)	2,057	2,496	1.51 (1.41-1.63)
AA	2,064	2,430	1.26 (1.17-1.35)	2,495	3,588	1.22 (1.14-1.31)	735	886	1.54 (1.37-1.72)

[a] Non-to-occasional drinkers consume less than 1 gram of alcohol per day.

[b] Light-to-moderate drinkers consume 1-28 grams of alcohol per day.

[c] Heavy drinkers consume more than 28 grams of alcohol per day.

[d] Adjusted for age, sex, study site, total energy intake, and the first three principal components.

[e] Non-drinking cases: GG (39%), GA (46%), AA (15%); Non-drinking controls: GG (37%), GA (47%), AA (16%).

[f] Light-to-moderate drinking cases: GG (34%), GA (48%), AA (18%); Light-to-moderate drinking controls: GG (36%), GA (47%), AA (17%).

[g] Heavy drinking cases: GG (32%), GA (50%), AA (18%); Heavy drinking controls: GG (32%), GA (50%), AA (18%).

Abbreviations: No. = Number.