Rare Variants in the DNA Repair Pathway and the Risk of Colorectal Cancer

Marco Matejcic[1], Hiba A. Shaban[1], Melanie W. Quintana[2], Fredrick R. Schumacher[3,4], Christopher K. Edlund[5], Leah Naghi[6], Rish K. Pai[7], Robert W. Haile[8], A. Joan Levine[8], Daniel D. Buchanan[9,10,11], Mark A. Jenkins[12], Jane C. Figueiredo[13], Gad Rennert[14], Stephen B. Gruber[15], Li Li[16], Graham Casey[17], David V. Conti[18], and Stephanie L. Schmit[1,19]

ABSTRACT

Background: Inherited susceptibility is an important contributor to colorectal cancer risk, and rare variants in key genes or pathways could account in part for the missing proportion of colorectal cancer heritability.

Methods: We conducted an exome-wide association study including 2,327 cases and 2,966 controls of European ancestry from three large epidemiologic studies. Single variant associations were tested using logistic regression models, adjusting for appropriate study-specific covariates. In addition, we examined the aggregate effects of rare coding variation at the gene and pathway levels using Bayesian model uncertainty techniques.

Results: In an exome-wide gene-level analysis, we identified *ST6GALNAC2* as the top associated gene based on the Bayesian risk index (BRI) method [summary Bayes factor $(BF)_{BRI} = 2604.23$]. A rare coding variant in this gene, rs139401613, was the top associated variant ($P = 1.01 \times 10^{-6}$) in an exome-wide single variant analysis. Pathway-level association analyses based on the integrative BRI (iBRI) method found extreme evidence of association with the DNA repair pathway ($BF_{iBRI} = 17852.4$), specifically with the nonhomologous end joining ($BF_{iBRI} = 437.95$) and nucleotide excision repair ($BF_{iBRI} = 36.96$) subpathways. The iBRI method also identified *RPA2, PRKDC, ERCC5*, and *ERCC8* as the top associated DNA repair genes (summary $BF_{iBRI} \geq 10$), with rs28988897, rs8178232, rs141369732, and rs201642761 being the most likely associated variants in these genes, respectively.

Conclusions: We identified novel variants and genes associated with colorectal cancer risk and provided additional evidence for a role of DNA repair in colorectal cancer tumorigenesis.

Impact: This study provides new insights into the genetic predisposition to colorectal cancer, which has potential for translation into improved risk prediction.

Introduction

Inherited genetic factors play an important role in the etiology of colorectal cancer (1). Genome-wide association studies (GWAS) in colorectal cancer have identified approximately 140 common risk loci predominantly in European and East Asian populations (2–4). However, risk variants, both high- and low-penetrance, collectively account for only a small proportion of the estimated familial risk (3). Part of the unexplained proportion of colorectal cancer heritability may be attributable to rare variants in genes/pathways with established or suspected roles in colorectal carcinogenesis, such as DNA repair, TGFβ signaling, vitamin D, and folate metabolism (5–8).

Rare variants by definition occur with a <1% frequency in the population and therefore large sample sizes are required to detect rare risk variants with modest effect sizes. Yet, next-generation sequencing approaches have enabled successful identification of population-specific rare variants and assessed their contributions to colorectal cancer risk. A recent exome sequencing study of early-onset colorectal cancer cases identified rare, highly penetrant pathogenic variants in 16% of cases, including novel variants in *POT1*, *POLE2*, and *MRE11* (9). Whole-genome sequencing of sporadic colorectal cancer cases identified a rare variant in CHD1 with a strong protective effect (2). Studies using exome-wide genotyping arrays have also identified rare risk variants with large effect sizes [odds ratio

[1]Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida. [2]Berry Consultants, Austin, Texas. [3]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio. [4]Seidman Cancer Center, University Hospitals, Cleveland, Ohio. [5]Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California. [6]Department of Medicine, Montefiore Medical Center, Albert Einstein College of Medicine, New York, New York. [7]Department of Laboratory Medicine and Pathology, Mayo Clinic Arizona, Scottsdale, Arizona. [8]Department of Medicine, Research Center for Health Equity, Cedars-Sinai Samuel Oschin Comprehensive Cancer Center, Los Angeles, California. [9]Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of Melbourne, Parkville, Victoria, Australia. [10]Victorian Comprehensive Cancer Centre, University of Melbourne, Centre for Cancer Research, Parkville, Victoria, Australia. [11]Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Parkville, Victoria, Australia. [12]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia. [13]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California. [14]Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel. [15]Center for Precision Medicine, City of Hope, Duarte, California. [16]Department of Family Medicine, University of Virginia, Charlottesville, Virginia. [17]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia. [18]Department of Preventive Medicine, Division of Biostatistics, University of Southern California, Los Angeles, California. [19]Department of Gastrointestinal Oncology, Moffitt Cancer Center, Tampa, Florida.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

D.V. Conti and S.L. Schmit contributed equally to this article.

Corresponding Author: Stephanie L. Schmit, Genomic Medicine Institute, Cleveland Clinic, 9500 Euclid Avenue/NE50, Cleveland, OH 44195. Phone: 216-444-3173; Fax: 216-636-1609; E-mail: schmits3@ccf.org

Cancer Epidemiol Biomarkers Prev 2021;30:895–903

doi: 10.1158/1055-9965.EPI-20-1457

(OR) ≥ 2.0] in genes previously unreported to be associated with colorectal cancer risk (i.e., *TCF7L2, RAB11FIP5, COL27A1*; refs. 10, 11). Recently, Bayesian model uncertainty techniques have been developed to maximize the power of rare variant association studies by incorporating uncertainty of both the inclusion of variants and the direction of association of each included variant in a model (12).

To further explore the contribution of rare coding variation to the risk of colorectal cancer, we examined the association between rare coding variants and colorectal cancer risk using participants of European ancestry from three large epidemiologic studies for a total of 5,293 individuals (2,327 cases and 2,966 controls). In addition to single variant testing, we investigated the aggregate effects of rare coding variation at the gene and pathway levels using the integrative Bayesian risk index (iBRI) method that incorporates external biological information to help refine the variant set selection procedure and ultimately localize the variant(s) most likely driving the association signal.

Materials and Methods

Study subjects

Study participants were individuals of European descent from North America, Europe, Australia, and Israel recruited from the Kentucky Case-Control Study (KY), the Colon Cancer Family Registry (CCFR), and the Molecular Epidemiology of Colorectal Cancer Study (MECC). Details on the study design and data collection for each study are provided in the Supplementary Materials and Methods. All cases had histologically confirmed adenocarcinoma of the colon or rectum. All participants provided written informed consent, and specimens and data were collected according to protocols approved by the Institutional Review Boards of the respective institutions.

Genotyping and quality control

Genomic DNA was extracted from blood samples and genotyped using the Infinium Human Exome BeadChip 12v1.0 (Illumina Inc.) plus custom content, which provides comprehensive coverage of >250,000 genetic markers across >20,000 genes in the human coding genome. Genotype calling was performed using standard methods.

Data from KY, CCFR, and MECC were cleaned on the basis of standard quality control (QC) metrics at the sample and variant levels as detailed in the Supplementary Materials and Methods. Among the 5,418 KY and CCFR samples genotyped, 2,691 KY samples (1,054 colorectal cancer cases and 1,637 controls) and 2,311 CCFR samples (1,132 colorectal cancer cases and 1,179 controls) passed QC filters and were retained for downstream analysis. Of the initial 1,392 MECC samples genotyped, 1,105 (559 colorectal cancer cases and 546 controls) passed QC filters, of which 303 were Arabs, 490 were Sephardi Jews, 291 were Ashkenazi Jews, and 21 were of unknown ethnicity. Only participants of Ashkenazi Jewish heritage (141 cases and 150 controls) were retained for downstream analyses due to their descent from Eastern Europe. Basic characteristics of the study subjects by case–control status are presented in Supplementary Table S1.

Overall, 142,390 polymorphic variants in 5,293 samples (2,327 cases and 2,966 controls) passed QC filters in at least one of the three studies and were assessed for single-variant association with colorectal cancer risk. A higher proportion of rare variants [minor allele frequency (MAF) <1%] was observed in KY (34.4%) and CCFR (33.2%) compared with MECC (10.1%; Supplementary Table S2).

For the gene- and pathway-level analyses, common variants (MAF ≥ 1%, $n = 24{,}425$) and variants in downstream ($n = 251$), intergenic ($n = 10{,}402$) and upstream ($n = 309$) regions based on the GRCh37 assembly were excluded. We considered a variant to be rare or common based on the 1% MAF cut-off in the study-specific samples that passed QC. ANNOVAR (2019Oct24 release) was used to annotate variants according to location and predicted functional effect (13). Potential pathogenic effects of nonsynonymous variants were predicted using ClinVar and six *in silico* algorithms [Combined Annotation-Dependent Depletion (CADD), Polyphen2-HumDiv (HDIV), PolyPhen2-HumVar (HVAR), Likelihood Ratio Test (LRT), Mutation Taster (MT), and Sorting Tolerant From Intolerant (SIFT)] from dbNSFP (14). Top associated variants were analyzed for potential functional or regulatory effects using established tools and resources such as HaploReg v4.1 (https://pubs.broadinstitute.org/mammals/haploreg), GTEx (https://www.gtexportal.org), RegulomeDB 2.0 (https://regulomedb.org), ENCODE (https://www.encodeproject.org/), and FuncPred (http://www.funcpred.com/).

Single variant association testing and annotation

The associations between allele dosages for all common (MAF ≥ 1%) and rare (<1%) variants and the risk of colorectal cancer were estimated through a 1-degree of freedom likelihood ratio test (LRT) assuming a log-additive model. Per-allele ORs and 95% confidence intervals (CI) were also estimated using unconditional logistic regression adjusting for appropriate study-specific covariates and principal components (PC) that capture sub-European ancestry. For KY and MECC, models were adjusted for age, sex, and the top four PCs. For CCFR, models were adjusted for age, sex, the top four PCs, and recruitment site (Australia, FHCRC, Mayo, Ontario, and USC). To summarize evidence of association across the three studies, we performed a meta-analysis from the sample size-weighted average of the study-specific test statistics using METAL (15). Exome-wide statistical significance was set at Bonferroni adjusted $P < 3.51 \times 10^{-7}$ in two-sided test based on the 142,390 unique polymorphic variants tested across

Table 1. Summary of iBRI results for association between the biological pathways investigated and colorectal cancer risk.

		KY (1,054 cases; 1,637 controls)			CCFR (1,132 cases; 1,179 controls)			MECC (141 cases; 150 controls)		
Pathway	Summary BF_{iBRI}[a]	Genes[b]	Rare variants[c]	BF_{iBRI}[d]	Genes	Variants	BF_{iBRI}	Genes	Variants	BF_{iBRI}
DNA repair	17,852.4	132	1,082	21.15	132	1,042	34.13	108	291	24.73
Folate metabolism	0.46	38	195	0.54	37	172	0.40	24	39	2.16
TGFβ signaling	0.37	75	280	0.96	72	301	0.50	36	67	0.77
Vitamin D metabolism	0.35	10	128	0.70	9	130	0.51	6	39	1.00
Total		255	1,685		250	1,645		174	436	

[a]Summary BF from the iBRI method calculated by the product of the three study-specific BFs (KY, CCFR, MECC).
[b]Number of genes in the pathway.
[c]Minor allele frequency <0.01.
[d]Study-specific BF from the iBRI method.

Table 2. Summary of iBRI results for association between DNA repair subpathways and colorectal cancer risk.

		KY (1,054 cases; 1,637 controls)			CCFR (1,132 cases; 1,179 controls)			MECC (141 cases; 150 controls)		
Subpathway	Summary BF_{iBRI}[a]	Genes[b]	Variants[c]	BF_{iBRI}[d]	Genes	Variants	BF_{iBRI}	Genes	Variants	BF_{iBRI}
ATM	0.23	15	140	0.14	15	136	0.05	12	41	28.76
BER	0.04	24	131	0.39	24	133	0.09	17	30	1.03
FA/HR	0.67	32	270	0.21	34	270	20.13	30	78	0.16
MMR	0.32	9	118	1.33	9	101	0.12	8	25	2.04
NER	36.96	15	109	11.59	12	101	4.37	10	27	0.73
NHEJ	437.85	9	63	121.94	9	61	9.26	6	9	0.39
OTHER	2.09	11	57	0.17	10	52	1.37	8	11	8.97
RECQ	0.1	4	58	0.18	4	58	0.32	4	24	1.72
TLS	1.95	9	82	0.36	11	76	4.3	10	27	1.28
XLR	1.01	4	54	1.2	4	54	1.11	3	19	0.75

[a]Summary BF from the iBRI method calculated by the product of the three study-specific BFs (KY, CCFR, MECC).
[b]Number of genes in the subpathway.
[c]Number of variants in the subpathway.
[d]Study-specific BF from the iBRI method.

the three studies. $P < 0.05$ and $P < 0.001$ were used to identify variants with nominal evidence of association.

We created quantile–quantile (Q–Q) plots of LRT *P* values and computed the genomic control factor (λ) to find evidence of residual population stratification after adjustment for PCs. Manhattan plots were generated to provide a visual illustration of the most significantly associated variants throughout the exome. Study-specific Q–Q plots and Manhattan plots were created using all variants and common variants only. The Q–Q plots including both common and rare variants showed evidence of inflation in both the study-specific *P* values ($\lambda = 1.54$–1.80; Supplementary Table S2) and meta *P* values ($\lambda = 1.67$; Supplementary Fig. S1A), which is likely due to the high sensitivity of the LRT to rare variants (16). When only common variants were plotted, a more uniform distribution between expected and observed *P* values was observed, with the genomic inflation factor indicating no appreciable evidence of population stratification after adjustment for PCs ($\lambda = 0.99$–1.07 for study-specific *P* values and $\lambda = 1.02$ for meta *P* values; Supplementary Fig. S1B). The Manhattan plots of meta-analysis estimates depict the magnitude of association of all variants (Supplementary Fig. S2A) and common variants only (Supplementary Fig. S2B) with colorectal cancer risk across the three studies (Supplementary Fig. S2).

Gene-level testing

The aggregate association of rare variants (MAF < 1%) with colorectal cancer risk was evaluated across gene-based regions, with variants assigned to genes based on chromosome position. Five methods were used for gene-level testing of rare variants: (i) weighted sum statistic by Madsen and Browning (17); (ii) sequence kernel association test (SKAT; ref. 18); (iii) computational step-up (19); and (iv) Bayesian Risk Index (BRI), which formally incorporates uncertainty of both the inclusion of variants in a region as well as the direction of effect of each included variant via Bayesian model uncertainty techniques (12). The BRI method is used to calculate a Bayes factor (BF) that quantifies the evidence that at least one variant within the specified region is associated with colorectal cancer. Details of each statistical method are described in the Supplementary Materials and Methods. Study-specific association models were adjusted for sex, age, the top four PCs, and for CCFR only, recruitment site. Evidence of gene-level associations across the three studies was found using a sample size-weighted meta-analysis of study-specific test statistics using METAL (15). The significance threshold for gene-level association analysis was $P < 3.61 \times 10^{-6}$ in two-sided test based on the 13,866 unique genes with ≥2 rare variants in any of the three studies. $P < 0.05$ was used to identify variants with nominal evidence of association. For the BRI approach, we calculated the product of the three study-specific BFs to obtain a summary BF under the assumption that studies are independent, each with its own study-specific prior probability of the null hypothesis, consistent with the BRI model (20). Gene-level associations with colorectal cancer risk based on the BRI method were defined as extreme (summary BF ≥ 100), strong (summary BF ≥ 10), and moderate (summary BF ≥ 1) according to *a priori* categorization (21).

Pathway-level testing

The association between rare variants (MAF < 1%) and colorectal cancer risk was evaluated across gene sets grouped by candidate pathways chosen based on *a priori* evidence for their involvement in colorectal carcinogenesis: DNA repair ($n_{gene} = 137$), TGFβ signaling ($n_{gene} = 88$), vitamin D metabolism ($n_{gene} = 10$), and folate metabolism ($n_{gene} = 42$; refs. 2, 5, 8). We further integrated information about specific DNA repair subpathways such as ataxia telangiectasia mutated (ATM), base excision repair (BER), Fanconi anemia/homologous recombination (HR/FA), mismatch repair (MMR), nucleotide excision repair (NER), nonhomologous end joining (NHEJ), recQ helicases (RECQ), translesion synthesis (TLS), cross-link repair (XLR), and other minor pathways (OTHER). DNA repair genes were annotated according to previous curations (22, 23). For TGFβ signaling, vitamin D metabolism and folate metabolism, genes were downloaded from KEGG version 91 (https://www.genome.jp/kegg/, release 2019/10) and REACTOME version 70 (https://reactome.org/, release 2019/09). A full list of genes assigned to each pathway as well as the number of common and rare variants identified in each study are summarized in Supplementary Table S3.

The pathway-level analysis was performed using the iBRI approach, which extends over the BRI method by integrating multiple region-specific risk indices within a given model and including external predictor-level covariates to help guide the region and variant selection procedure (24). Details of the iBRI method are described in the Supplementary Materials and Methods. Briefly, we used pathway information to estimate the probability that at least one variant within the specified pathway is associated with colorectal cancer risk through estimation of the BF via set-specific posterior probability, which is the sum of the posterior model probabilities for every model that includes

at least one variant within the set (12). The iBRI was computed separately for each gene in the pathway, and multiple gene-specific risk indices were included in a single model adjusted for sex, age, the top four PCs, and for CCFR only, recruitment site. Given evidence of an association with the DNA repair pathway, we further investigated the likely genes and variants that are driving the pathway-level association through estimation of BFs under the iBRI method. For each study, the top 10 DNA repair genes within the top 25 models found using the iBRI method were plotted. Summary BFs were calculated as the product of the three study-specific BFs, using a weighted average of the posterior probabilities of inclusion for the three studies (20). Similarly to the BRI method, strength of association with colorectal cancer risk based on iBRI was defined on *a priori* categorization of the summary BF (21).

All statistical analyses were conducted using the R statistical computing platform and PLINK v. 1.07 (25).

Results

A flow diagram representing the analysis pipeline and key results from the study is provided in **Fig. 1**.

Single variant association analysis

The exome-wide single variant analysis was carried out on 142,390 polymorphic variants that passed QC in at least one of the three studies investigated. Although none of the variants reached exome-wide statistical significance, 7,099 variants were nominally associated with colorectal cancer risk (meta $P < 0.05$), including 84 variants at the 0.001 significance level. These 84 variants with meta $P < 0.001$ are summarized in Supplementary Table S4. Of these, 56 (66.7%) were rare or monomorphic among controls in the overall study population. The strongest association was found for three low-frequency variants located on chromosome 17: rs139401613 in *ST6GALNAC2* (meta $P = 1.01 \times 10^{-6}$; 0.034% in overall controls), rs35467001 (meta OR = 2.05; 95% CI, 1.5–2.81; meta $P = 7.72 \times 10^{-6}$; 3.9% in overall controls) and rs34322745 (meta OR = 2.1; 95% CI, 1.47–3.02; meta $P = 5.16 \times 10^{-5}$; 3.6% in overall controls) in *SDK2*. A noteworthy association was also observed for rs61736607 in *ADAMTS14* (meta OR = 1.67; 95% CI, 1.3–2.14; meta $P = 7.1 \times 10^{-5}$; 7.3% in overall controls). While coding variants rs35467001 and rs34322745 in *SDK2* were in substantial LD across the three studies ($r^2 = 0.74$–0.79), none of these were in LD with rs139401613 ($r^2 \leq 0.00052$). The two top associated variants (rs139401613 and rs35467001) were each predicted to be deleterious based on three out of the six algorithms examined.

Gene-level association analysis

A total of 12,967 genes in KY, 12,729 genes in CCFR, and 5,582 genes in MECC harbored ≥2 rare variants and were assessed for association with colorectal cancer risk. On the basis of SKAT, none of the genes reached exome-wide statistical significance ($P < 3.61 \times 10^{-6}$), with the strongest associations observed for *SPTBN5* (4.66×10^{-5}), *CAVIN1* (9.88×10^{-5}), and *ST6GALNAC2* (1.1×10^{-4}). Using the BRI approach, 17 genes were extremely associated (summary $BF_{BRI} \geq 100$), 300 genes were strongly associated (summary $BF_{BRI} \geq 10$), and 3,744 genes were moderately associated (summary $BF_{BRI} \geq 1$) with colorectal cancer risk. The top associated genes with summary $BF_{BRI} \geq 100$ are presented in Supplementary Table S5. The strongest association was found for *ST6GALNAC2* on chromosome 17 (summary $BF_{BRI} = 2604.23$), followed by *OSTM1, COL22A1, EPHA7, TTC28, SPTBN5, FSIP1, AKR1D1, NOTCH3, C6orf120, OR11H4, NAT1, EVI2B, CENPQ, SMPDL3A, CEP43*, and *GPC3* with summary BF_{BRI} ranging from 104.43 to 948.28. All these genes were nominally associated with colorectal cancer risk (meta $P < 0.05$) in at least one of the alternative gene-level tests (weighted sum, SKAT, and computational step-up), except for *FSIP1* that showed a borderline association in the weighted sum test (meta $P = 0.054$).

Pathway-level analysis

Table 1 summarizes the number of genes and rare variants included in each of the pathways investigated as well as the pathway-level association results based on the iBRI method. The study-specific BF_{iBRI} ranges were 21.15 to 34.13 for DNA repair pathway, 0.50 to 0.96 for TGFβ signaling, 0.51 to 1.00 for vitamin D metabolism, and 0.40 to 2.16 for folate metabolism. The summary BF from the iBRI method highlighted an extreme association between DNA repair pathway and colorectal cancer risk (summary $BF_{iBRI} = 17852.4$). **Figure 2** plots the study-specific inclusion of the top 10 DNA repair genes in the top 25 models found using iBRI. Given evidence of association with the DNA repair pathway, we further estimated the risk associated with individual DNA repair subpathways under the iBRI method (**Table 2**). We found an extreme association with NHEJ (summary $BF_{iBRI} = 437.85$) and a strong association with NER (summary $BF_{iBRI} = 36.96$). The other DNA repair subpathways were reported with $BF_{iBRI} < 10$ in either the summary or study-specific results.

The iBRI method also identified the most likely genes and variants driving the association with DNA repair pathway. **Table 3** reports genes with summary $BF_{iBRI} \geq 1$ and, for each gene, the variant with the highest summary BF_{iBRI} (defined as the most likely associated variant). The top associated gene was *RPA2* (summary $BF_{iBRI} = 164.28$) in the group of genes representing minor DNA repair pathways (OTHER), and the most likely associated variant in this gene was rs28988897 (summary $BF_{iBRI} = 657.56$). An extreme association was also found for *PRKDC* (summary $BF_{iBRI} = 109.77$) in NHEJ, with the most likely associated variant being rs8178232 (summary $BF_{iBRI} = 17.68$). Strong associations were observed for *ERCC5* (summary $BF_{iBRI} = 65.52$) and *ERCC8* (summary $BF_{iBRI} = 21.9$) in NER. The most likely associated variants in these genes were rs141369732 (summary $BF_{iBRI} = 6.39$) and rs201642761 (summary $BF_{iBRI} = 0.92$), respectively. Additional information on the DNA repair genes and variants associated with colorectal cancer risk based on the iBRI method are provided in Supplementary Tables S6 and S7. *RPA2, ERCC5, ERCC8, XPC, LIG4*, and *MSH3* were nominally associated with colorectal cancer risk (meta $P < 0.05$) in at least one of the alternative gene-level tests (weighted sum, SKAT, and computational step-up), with *XPC* being the only gene consistently associated across all tests. In addition, there was a substantial increase in summary BF estimates from the basic BRI approach to the iBRI method for the top associated genes (*RPA2, PRKDC, ERCC5*, and *ERCC8*; summary $BF_{iBRI} = 21.9$–164.28 vs. summary $BF_{BRI} = 2.038$–6.53; Supplementary Table S6).

A sensitivity analysis excluding MECC subjects showed robust associations only for the DNA repair pathway (summary $BF_{iBRI} = 721.85$) and the NER and NHEJ subpathways (summary $BF_{iBRI} = 50.65$ and 1129.16, respectively), indicating that a limited sample size is not driving the lack of association with other pathways. At the gene-level, there was variability in summary BF when the association was mainly driven by MECC, with *RPA2* being no longer strongly associated with colorectal cancer risk ($BF_{iBRI} = 2.2$). This heterogeneity in results across cohorts reinforces the hypothesis that population-specific rare variants are driving the association signal within the same gene.

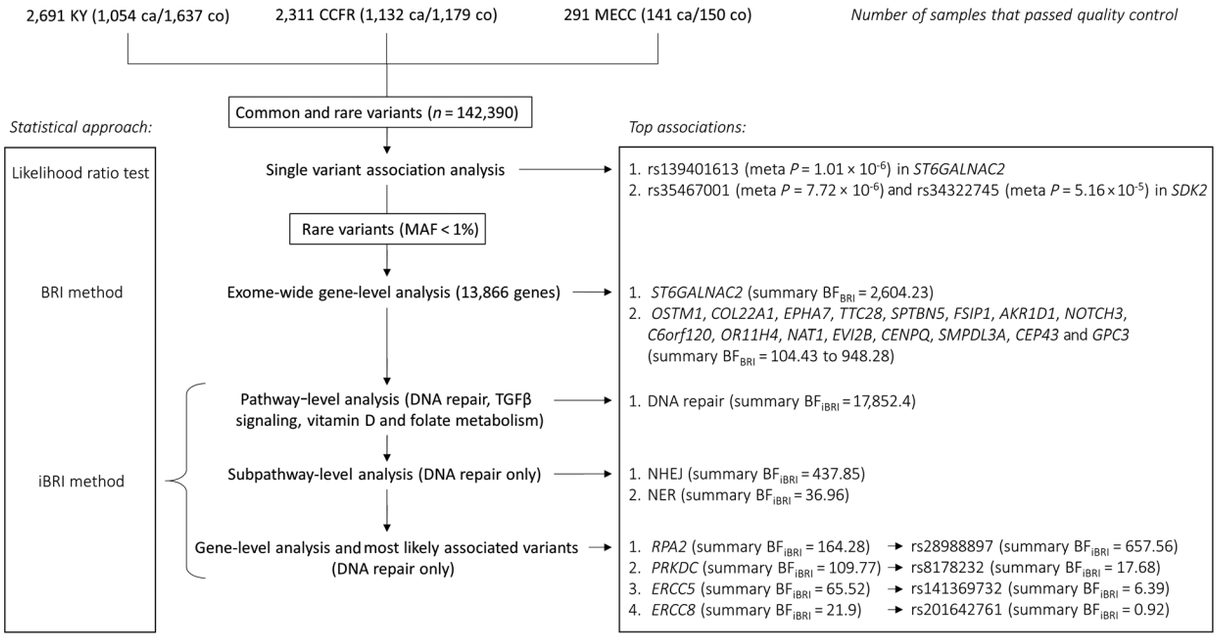

Figure 1.

Project flow chart summarizing analyses and key results. A total of 142,390 polymorphic variants in 5,293 samples (2,327 colorectal cancer cases and 2,966 controls) were analyzed. The associations between common and rare variants and the risk of colorectal cancer were estimated through the likelihood ratio tests. A meta-analysis of the study-specific test statistics revealed strong associations between colorectal cancer and rs139401613 in *ST6GALNAC2* (meta $P = 1.01 \times 10^{-6}$), rs35467001 (meta $P = 7.72 \times 10^{-6}$; 3.9%), and rs34322745 (meta $P = 5.16 \times 10^{-5}$) in *SDK2*. Only rare variants (MAF < 1%) were retained for gene- and pathway-level analyses. A summary BF was computed as the product of the three study-specific BFs from the BRI method. The top associated gene was *ST6GALNAC2* (summary $BF_{BRI} = 2{,}604.23$) followed by *OSTM1, COL22A1, EPHA7, TTC28, SPTBN5, FSIP1, AKR1D1, NOTCH3, C6orf120, OR11H4, NAT1, EVI2B, CENPQ, SMPDL3A, CEP43*, and *GPC3* with summary BF_{BRI} ranging from 104.43 to 948.28. The iBRI method was used to perform pathway-level analysis for DNA repair, TGFβ signaling, and vitamin D and folate metabolism. Given evidence of strong association with the DNA repair pathway (summary $BF_{BRI} = 17{,}852.4$), we further investigated the likely subpathways, genes, and variants that are driving the association. At the subpathway level, extreme associations were found for NHEJ (summary $BF_{iBRI} = 437.85$) and NER (summary $BF_{iBRI} = 36.96$). The top associated gene was *RPA2* (summary $BF_{iBRI} = 164.28$), and the most likely associated variant in this gene was rs28988897 (summary $BF_{iBRI} = 657.56$). Strong associations were also reported for *PRKDC* (summary $BF_{iBRI} = 109.77$) [rs8178232 (summary $BF_{iBRI} = 17.68$)], *ERCC5* (summary $BF_{iBRI} = 65.52$) [rs141369732 (summary $BF_{iBRI} = 6.39$)], and *ERCC8* (summary $BF_{iBRI} = 21.9$) [rs201642761 (summary $BF_{iBRI} = 0.92$)]. Abbreviations: ca, cases; co, controls; meta P = meta-analysis P value.

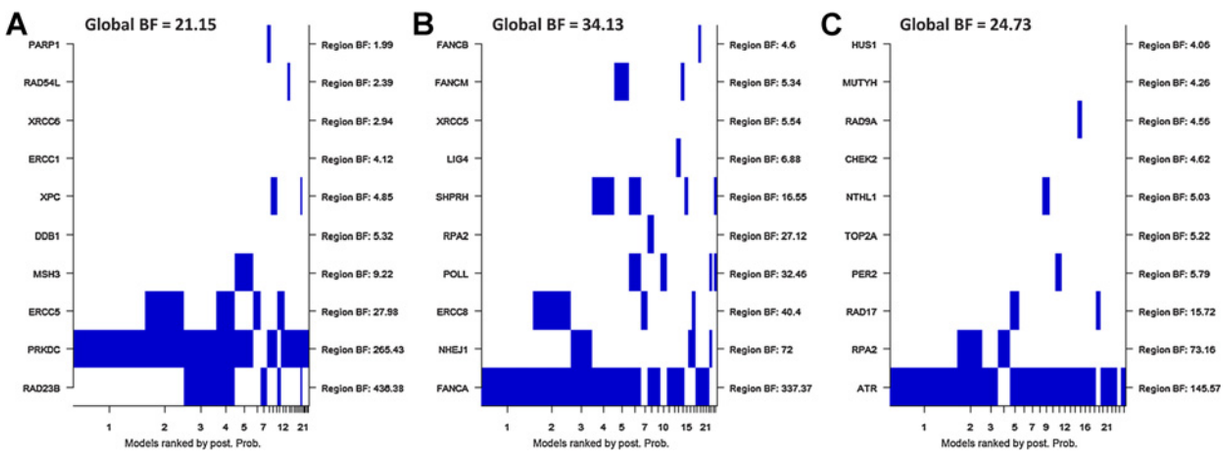

Figure 2.

Top model inclusions for top DNA repair genes. Top 10 DNA repair genes included in the top 25 models identified using the iBRI approach in each study: KY (**A**), CCFR (**B**), and MECC (**C**). The top 10 genes ordered by iBRI BF are plotted on the left axis, and the respective iBRI BFs are reported on the right. The top 25 models ordered by posterior probability are plotted on the *x*-axis. Within the plot, each blue rectangle represents the inclusion of a gene within the respective model, and the width of each column is proportional to the posterior model probability. A gene is defined as being included in a model if at least one variant within the region was included in the model.

Table 3. Summary of iBRI results for the top associated DNA repair genes and most likely associated variant in each gene.

Gene-level associations[a]						Variant-level associations[a]						
			Study-specific BF_{iBRI}[b]							Study-specific BF_{iBRI}		
Gene[c]	Subpathway	Summary BF_{iBRI}[d]	KY	CCFR	MECC	Total variants[e]	Most likely associated variant[f]	dbSNP[g]	Summary BF_{iBRI}	KY	CCFR	MECC
RPA2	OTHER	164.28	0.08	27.12	73.16	2	exm37202	rs28988897	657.56	0.17	54.26	73.16
PRKDC	NHEJ	109.77	265.43	1.74	0.24	39	exm699630	rs8178232	17.68		17.68	
ERCC5	NER	65.52	27.98	1.05	2.22	23	exm1078156	rs141369732	6.39			6.39
ERCC8	NER	21.9	0.54	40.4		8	exm456700	rs201642761	0.92		0.92	
REV3L	TLS	3.57	0.98	1.71	2.13	30	exm572168	rs189752287	32.14		32.14	
XPC	NER	3.39	4.85	4.15	0.17	13	exm292542	rs121965090	46.91	18.56	2.53	
LIG4	NHEJ	2.88	1.79	6.88	0.23	12	exm1078631	rs201683262	3.48	3.28	1.06	
MSH3	MMR	2.42	9.22	0.37	0.7	15	exm464610	rs35009542	2.34	2.34		
POLQ	XLR	1.9	0.86	2.11	1.04	36	custom_3.121212545	rs767282392	0.75		0.75	
TDG	BER	1.41			1.41	1	exm1031747	rs140436257	1.41			1.41

[a]See Supplementary Tables S6 and S7 for details on gene- and variant-level associations.
[b]Study-specific BF from the iBRI method; empty cells if no rare variants in the gene.
[c]Only genes with summary BF ≥ 1 from the iBRI method are shown; genes are ordered by decreasing summary BF.
[d]Summary BF from the iBRI method.
[e]Total number of unique rare variants identified in the gene across the three studies (KY, CCFR, MECC).
[f]Most likely associated variant in the gene based on summary BF from the iBRI method.
[g]dbSNP identifiers retrieved from the National Center for Biotechnology Information (NCBI).

Discussion

This study examined the association between rare coding variants and colorectal cancer risk using Bayesian approaches that maximize the power to detect rare variant associations at multiple levels (i.e., pathway, gene and variant levels). The more efficient model search algorithm and the inclusion of external information (i.e., biological pathway) to help guide the variant set selection procedure allowed the identification of novel rare risk variants and provided additional evidence for a role of DNA repair pathways and genes in colorectal cancer tumorigenesis.

Our analysis based on the iBRI method identified NHEJ and NER as the only DNA repair subpathways strongly associated with colorectal cancer risk. These findings are consistent with the well-recognized role of these pathways in carcinogenesis as well as with previous studies reporting associations between genetic polymorphisms in genes involved in these pathways and colorectal cancer risk (26, 27). Using DNA repair subpathways as predictor-level covariates to help estimate the likelihood that any variant in a gene is associated with colorectal cancer risk, we found extreme gene-level associations with *RPA2* and *PRKDC*. *RPA2* encodes for a subunit of the heterotrimeric replication protein A (RPA) complex that is essential for DNA replication, repair, recombination, and cell-cycle regulation (28). *PRKDC* in the NHEJ pathway encodes for a nuclear serine/threonine protein kinase that plays a pivotal role in DNA damage response and maintenance of genomic stability (29). Aberrant expression of *RPA2* and *PRKDC* has been previously associated with colorectal cancer development and adverse clinical outcome (30, 31). We also found strong evidence of association for *ERCC5* and *ERCC8*, which are essential components of the NER pathway. Genetic polymorphisms in *ERCC5* have been previously associated with risk and prognosis of colorectal cancer (32). *ERCC8* has been identified as a potential susceptibility factor for certain cancers (33); however, our study is the first to report an association specifically with colorectal cancer. Among the other DNA repair genes with moderate evidence of association with colorectal cancer risk, *REV3L, XPC, MSH3, POLQ*, and *TDG* have been previously associated with either familial (34) and sporadic colorectal cancer (35, 36). Although many of these genes are recognized tumor suppressor genes or oncogenes, further work is required to investigate their potential role as genetic biomarkers and therapeutic targets for colorectal cancer.

In addition to gene-level associations, the iBRI method allowed to identify the most likely associated variant in each of the top associated genes. There is no prior evidence of association between these variants and colorectal cancer risk in the literature. However, except for rs121965090 in *XPC*, all these variants were predicted to be deleterious according to at least one of the algorithms examined. On the basis of ENCODE data, the most likely associated variants in *RPA2*, rs28988897, overlaps with multiple histone marks associated with enhancer and promoter activity in colon and rectal samples. In addition, this variant is likely to affect binding of transcription factors according to RegulomeDB (probability score >0.8), supporting its role as regulatory variant. These findings demonstrate the usefulness of predictor-level covariates such as biological pathways in identifying the variants that are most likely driving the associations observed at the gene level and that would not be detected by other statistical methods.

In the exome-wide gene-level analysis, *ST6GALNAC2* (17q25.1) emerged as the strongest association based on the BRI method. *ST6GALNAC2* encodes for a type II transmembrane Golgi-localized enzyme involved in protein glycosylation, which is a key post-translational modification that plays a role in regulating multiple cellular processes including cell adhesion, migration, signaling and immune response (37). Aberrant protein glycosylation is a common phenotypic alteration occurring in many types of cancer, including colorectal cancer (38). A rare variant located in exon 9 of *ST6GALNAC2*, rs139401613, was the top associated variant in the single variant analysis. This nonsynonymous variant results in the Arg340Gln amino acid change that was predicted to be deleterious by SIFT, HDIV, and HVAR. On the basis of ENCODE data specific to colon and rectal samples, rs139401613 overlaps with histone marks that are indicative of enhancer and promoter activity. In addition, this variant might affect binding of transcription factors and influence gene expression through enhancer activity.

Among the other genes with strong evidence of association from the BRI method, *EPHA7*, *TTC28*, *EVI2B, GPC3, CENPQ, NOTCH3*, and *SMPDL3A* were previously reported to play a role in the development and prognosis of colorectal cancer (39–45). Our study is the first to report an association of *C6orf120*, *COL22A1, FSIP1, AKR1D1*, and *CEP43* specifically with colorectal cancer risk. Consistency in the associations between these genes and colorectal cancer risk in at least one of the alternative gene-level tests (weighted sum, SKAT, and computational step-up) increased our confidence in the validity of the results. These genes encode for proteins involved in signal transduction and cell differentiation (*GPC3, EPHA7, NOTCH3*; refs. 46–48), immunoregulation (C6orf120; ref. 49), structural and physiological integrity (*COL22A1*; ref. 50), cell-cycle progression and growth regulation (*TTC28, CEP43, EVI2B, FSIP1, CENPQ*; refs. 51–55), bile acid synthesis and steroid hormone metabolism (*AKR1D1*; ref. 56), and oxysterol or lipid metabolism *SMPDL3A* (57). Thus, our findings not only confirm the importance of established colorectal cancer-related genes, but also highlight novel genes previously unreported to be involved in colorectal cancer tumorigenesis.

In the exome-wide single variant analysis, 84 variants were associated with colorectal cancer risk at the meta $P < 0.001$ significance level. Of the 7,099 variants nominally associated with colorectal cancer risk (meta $P < 0.05$), 13 common variants were previously reported to be associated with colorectal cancer risk in previous exome- or genome-wide association studies (Supplementary Table S8; refs. 2, 3, 58). On the other hand, we failed to replicate associations with rare colorectal cancer risk variants previously identified in populations of European descent (2, 9, 11, 58), except for two variants (rs2427284 and rs13042941) that were previously associated with colorectal cancer risk in conditional analyses (Supplementary Table S8; ref. 58). There are several factors that may have hampered the replication of previous findings. First, the majority of prior studies were conducted using familial colorectal cancer cases that are enriched for rare high-risk alleles as compared with sporadic cases. Second, rare risk variants are usually population-specific and associations with these variants may not be replicated across different ancestral groups (59). Third, it is possible that previous colorectal cancer studies were underpowered for detecting associations with low-penetrance, rare risk variants. Finally, risk variants may interplay with other genetic and/or environmental factors that could influence their expressivity and pathogenic effect (60). Further studies using admixed populations, leveraging large sample sizes, and evaluating gene–gene and gene–environment interactions are warranted to gain new insights into the pathogenesis of colorectal cancer.

Substantial heterogeneity across study-specific BFs from the BRI and iBRI methods was observed in our study. As rare variants are often population-specific and the number of recruited participants varies widely across the study populations investigated (i.e., MECC has a far smaller sample size than KY or CCFR), it is reasonable to observe variability in rare variation patterns across these different European-descent populations that may lead to divergent risk estimates. It is also possible that environmental or lifestyle factors that vary across populations could modify the associations between genetic variation and risk of colorectal cancer. Caution should be taken when generalizing the predicted risk obtained from one study to the general European-descent population. Nonetheless, the summary BF calculated as the product of the three study-specific BFs highlights an overall magnitude of association that warrants further investigation.

Taken together, the observed gene-level associations with colorectal cancer risk based on the Bayesian approaches are biologically plausible. As shown in **Fig. 2**, many of the top 25 models from the iBRI method were comprised of risk indices that included at least one variant in multiple DNA repair genes. Thus, our results support the polygenic inheritance model where the risk of colorectal cancer is influenced by multiple low-penetrance genes rather than a few high-penetrance cancer-predisposition genes. This study is the largest exome-focused association analysis of colorectal cancer using array-based technology and takes advantage of a large sample size derived from the combination of three large epidemiologic studies. However, there are several limitations such as the use of exome array that does not provide whole exome coverage and could miss a substantial proportion of very rare genetic variation. We did not investigate large indels that could result in loss-of-function of the entire protein. Also, our study was limited to the investigation of the coding genome and we cannot exclude potentially important effects from rare regulatory variation in noncoding regions. As many of the controls had a family history, the meta OR for colorectal cancer could be biased (presumably downwards); however, the *P* value for significance is expected to be robust, thus preserving the ranking of variants. Finally, we were unable to assess the generalizability of our results through an appropriate external validation dataset, although the combined analysis of three independent studies was important to identify and replicate potential candidate risk variants and genes for future studies.

In conclusion, the increased power and more efficient model search algorithms to identify rare risk variants provided new insights into the genetic and biological landscape of colorectal cancer by identifying novel colorectal cancer genes involved in DNA repair mechanisms and other key cellular processes such as cell division, signaling and immune regulation. Future work should focus on replicating our findings in independent cohorts and using more comprehensive methodologies such as targeted sequencing or fine-mapping to identify potential novel genetic targets for improved risk prediction.

Authors' Disclosures

No disclosures were reported.

Authors' Contributions

M. Matejcic: Data curation, formal analysis, investigation, writing–original draft, writing–review and editing. **H.A. Shaban:** Data curation, formal analysis. **M.W. Quintana:** Methodology. **F.R. Schumacher:** Data curation, investigation. **C.K. Edlund:** Resources, data curation. **L. Naghi:** Resources, writing–review and editing. **R.K. Pai:** Data curation, writing–review and editing. **R.W. Haile:** Resources, data curation, writing–review and editing. **A.J. Levine:** Writing–review and editing. **D.D. Buchanan:** Resources, data curation. **M.A. Jenkins:** Resources, data curation, funding acquisition, writing–review and editing. **J.C. Figueiredo:** Writing–review and editing. **G. Rennert:** Resources, funding acquisition. **S.B. Gruber:** Writing–review and editing. **L. Li:** Resources, data curation, funding acquisition. **G. Casey:** Data curation, methodology, writing–review and editing. **D.V. Conti:** Conceptualization, supervision, validation, investigation, methodology, writing–review and editing. **S.L. Schmit:** Conceptualization, formal analysis, supervision, investigation, project administration, writing–review and editing.

Acknowledgments

The Colon Cancer Family Registry (CCFR) was supported by the NCI of the NIH under award number U01 CA167551. Additional support for case ascertainment was provided from the Surveillance, Epidemiology, and End Results (SEER) Program of the NCI and the following U.S. state cancer registries: AZ, CO, MN, NC, NH; and by the Victoria Cancer Registry (Australia) and Ontario Cancer Registry (Canada). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Colon CFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government, any cancer registry, or the Colon CFR. The Kentucky Case-Control Study (KY) was supported by Damon Runyon Cancer Research Foundation Clinical Investigator Award, CI-8 (to L. Li); the Case Center for Transdisciplinary Research on Energetics and Cancer, U54 CA-116867-01 (to

L. Li); NCI K22 Award, K22 CA120545-01 (to L. Li); State of Ohio Biomedical Research and Technology Transfer Commission; and the NCI Award R25 CA094186-06.

The Molecular Epidemiology of Colorectal Cancer Study (MECC) was supported by the NCI at the NIH: R01 CA197350, R01 CA81488, P30 CA014089, U19 CA148107 (to S.B. Gruber), and P01 CA196569 and a generous gift from Daniel and Maryann Fong. This work was also supported by the National Human Genome Research Institute at the NIH [T32 HG000040] and the National Institute of Environmental Health Sciences at the NIH [T32 ES013678].

The CCFR acknowledges the generous contributions of their study participants, dedication of study staff, and the financial support from the U.S. NCI, without which this important registry would not exist.

We would also like to acknowledge the recipients of the following grants: State of Ohio Biomedical Research and Technology Transfer Commission (to G. Casey), the NCI Award R25 CA094186-06 (to C.L. Thompson), P01 CA196569 (to D.C. Thomas), the National Human Genome Research Institute at the NIH (T32 HG000040, to M. Boehnke), and the National Institute of Environmental Health Sciences at the NIH (T32 ES013678, to W. Gauderman).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received October 8, 2020; revised December 14, 2020; accepted February 22, 2021; published first February 24, 2021.

References

1. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and familial colon cancer. Gastroenterology 2010;138:2044–58.
2. 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. Nat Genet 2019; 51:76–87.
3. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nat Commun 2019;10:2154.
4. Lu Y, Kweon SS, Tanikawa C, Jia WH, Xiang YB, Cai Q, et al. Large-scale genome-wide association study of east asians identifies loci associated with risk for colorectal cancer. Gastroenterology 2019;156:1455–66.
5. Dou R, Ng K, Giovannucci EL, Manson JE, Qian ZR, Ogino S. Vitamin D and colorectal cancer: molecular, epidemiological and clinical evidence. Br J Nutr 2016;115:1643–60.
6. Jung B, Staudacher JJ, Beauchamp D. Transforming growth factor β superfamily signaling in development of colorectal cancer. Gastroenterology 2017;152:36–52.
7. Liu J, Zheng B, Li Y, Yuan Y, Xing C. Genetic polymorphisms of DNA repair pathways in sporadic colorectal carcinogenesis. J Cancer 2019;10:1417–33.
8. Moazzen S, Dolatkhah R, Tabrizi JS, Shaarbafi J, Alizadeh BZ, de Bock GH, et al. Folic acid intake and folate status and colorectal cancer risk: A systematic review and meta-analysis. Clin Nutr 2018;37(6 Pt A):1926–34.
9. Chubb D, Broderick P, Dobbins SE, Frampton M, Kinnersley B, Penegar S, et al. Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. Nat Commun 2016;7:11883.
10. Chang J, Tian J, Yang Y, Zhong R, Li J, Zhai K, et al. A rare missense variant in TCF7L2 associates with colorectal cancer risk by interacting with a GWAS-identified regulatory variant in the MYC enhancer. Cancer Res 2018;78:5164–72.
11. Jiao X, Liu W, Mahdessian H, Bryant P, Ringdahl J, Timofeeva M, et al. Recurrent, low-frequency coding variants contributing to colorectal cancer in the Swedish population. PLoS One 2018;13:e0193547.
12. Quintana MA, Berstein JL, Thomas DC, Conti DV. Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. Genet Epidemiol 2011;35:638–49.
13. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.
14. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. Hum Mutat 2011;32:894–9.
15. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;26:2190–1.
16. Pirie A, Wood A, Lush M, Tyrer J, Pharoah PD. The effect of rare variants on inflation of the test statistics in case-control analyses. BMC Bioinformatics 2015; 16:53.
17. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLos Genet 2009;5:e1000384.
18. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 2011;89:82–93.
19. Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PLoS One 2010;5:e13584.
20. Ly A, Etz A, Marsman M, Wagenmakers EJ. Replication Bayes factors from evidence updating. Behavior Research Methods 2019;51:2498–508.
21. Schönbrodt FD, Wagenmakers EJ. Bayes factor design analysis: Planning for compelling evidence. Psychon Bull Rev 2018;25:128–42.
22. Kang J, D'Andrea AD, Kozono D. A DNA repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. J Natl Cancer Inst 2012;104:670–81.
23. Wood RD, Mitchell M, Lindahl T. Human DNA repair genes, 2005. Mutat Res 2005;577:275–83.
24. Quintana MA, Conti DV. Integrative variable selection via Bayesian model uncertainty. Stat Med 2013;32:4938–53.
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
26. Bau DT, Yang MD, Tsou YA, Lin SS, Wu CN, Hsieh HH, et al. Colorectal cancer and genetic polymorphism of DNA double-strand break repair gene XRCC4 in Taiwan. Anticancer Res 2010;30:2727–30.
27. Berndt SI, Platz EA, Fallin MD, Thuita LW, Hoffman SC, Helzlsouer KJ. Genetic variation in the nucleotide excision repair pathway and colorectal cancer risk. Cancer Epidemiol Biomarkers Prev 2006;15:2263–9.
28. Borgstahl GE, Brader K, Mosel A, Liu S, Kremmer E, Goettsch KA, et al. Interplay of DNA damage and cell cycle signaling at the level of human replication protein A. DNA Repair (Amst) 2014;21:12–23.
29. Goodwin JF, Knudsen KE. Beyond DNA repair: DNA-PK function in cancer. Cancer Discov 2014;4:1126–39.
30. Givalos N, Gakiopoulou H, Skliri M, Bousboukea K, Konstantinidou AE, Korkolopoulou P, et al. Replication protein A is an independent prognostic indicator with potential therapeutic implications in colon cancer. Mod Pathol 2007;20:159–66.
31. Sun S, Cheng S, Zhu Y, Zhang P, Liu N, Xu T, et al. Identification of PRKDC (protein kinase, DNA-activated, catalytic polypeptide) as an essential gene for colorectal cancer (CRCs) cells. Gene 2016;584:90–6.
32. Li YK, Xu Q, Sun LP, Gong YH, Jing JJ, Xing CZ, et al. Nucleotide excision repair pathway gene polymorphisms are associated with risk and prognosis of colorectal cancer. World J Gastroenterol 2020;26:307–23.
33. Jing JJ, Sun LP, Xu Q, Yuan Y. Effect of ERCC8 tagSNPs and their association with H. pylori infection, smoking, and alcohol consumption on gastric cancer and atrophic gastritis risk. Tumour Biol 2015;36:9525–35.
34. Broderick P, Bagratuni T, Vijayakrishnan J, Lubbe S, Chandler I, Houlston RS. Evaluation of NTHL1, NEIL1, NEIL2, MPG, TDG, UNG and SMUG1 genes in familial colorectal cancer predisposition. BMC Cancer 2006;6:243.
35. Gil J, Ramsey D, Stembalska A, Karpinski P, Pesz KA, Laczmanska I, et al. The C/A polymorphism in intron 11 of the XPC gene plays a crucial role in the modulation of an individual's susceptibility to sporadic colorectal cancer. Mol Biol Rep 2012;39:527–34.
36. Jiraskova K, Hughes DJ, Brezina S, Gumpenberger T, Veskrnova V, Buchler T, et al. Functional polymorphisms in DNA repair genes are associated with sporadic colorectal cancer susceptibility and clinical outcome. Int J Mol Sci 2018;20:97.
37. Ohtsubo K, Marth JD. Glycosylation in cellular mechanisms of health and disease. Cell 2006;126:855–67.
38. Venkitachalam S, Revoredo L, Varadan V, Fecteau RE, Ravi L, Lutterbaugh J, et al. Biochemical and functional characterization of glycosylation-associated mutational landscapes in colon cancer. Sci Rep 2016;6:23642.
39. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7.

40. Foda AA, Mohammad MA, Abdel-Aziz A, El-Hawary AK. Relation of glypican-3 and E-cadherin expressions to clinicopathological features and prognosis of mucinous and non-mucinous colorectal adenocarcinoma. Tumour Biol 2015;36: 4671–9.
41. Guo YA, Chang MM, Huang W, Ooi WF, Xing M, Tan P, et al. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat Commun 2018;9:1520.
42. Pesson M, Volant A, Uguen A, Trillet K, De La Grange P, Aubry M, et al. A gene expression and pre-mRNA splicing signature that marks the adenoma-adenocarcinoma progression in colorectal cancer. PLoS One 2014;9:e87761.
43. Serafin V, Persano L, Moserle L, Esposito G, Ghisi M, Curtarello M, et al. Notch3 signalling promotes tumour growth in colorectal cancer. J Pathol 2011;224:448–60.
44. Wang J, Kataoka H, Suzuki M, Sato N, Nakamura R, Tao H, et al. Down-regulation of EphA7 by hypermethylation in colorectal cancer. Oncogene 2005; 24:5637–47.
45. Yuan Y, Chen J, Wang J, Xu M, Zhang Y, Sun P, et al. Identification hub genes in colorectal cancer by integrating weighted gene co-expression network analysis and clinical validation in vivo and vitro. Front Oncol 2020;10:638.
46. Brandstadter JD, Maillard I. Notch signalling in T cell homeostasis and differentiation. Open Biol 2019;9:190187.
47. Capurro MI, Xiang YY, Lobe C, Filmus J. Glypican-3 promotes the growth of hepatocellular carcinoma by stimulating canonical Wnt signaling. Cancer Res 2005;65:6245–54.
48. Li S, Wu Z, Ma P, Xu Y, Chen Y, Wang H, et al. Ligand-dependent EphA7 signaling inhibits prostate tumor growth and progression. Cell Death Dis 2017;8: e3122.
49. Li X, Qiao Y, Chang LS, Xiao F, Lu LH, Hao XH, et al. Role of C6ORF120, an N-glycosylated protein, is implicated in apoptosis of $CD4^{+}$ T lymphocytes. Chin Med J 2011;124:3560–7.
50. Koch M, Schulze J, Hansen U, Ashwodt T, Keene DR, Brunken WJ, et al. A novel marker of tissue junctions, collagen XXII. J Biol Chem 2004;279:22514–21.
51. Acquaviva C, Chevrier V, Chauvin JP, Fournier G, Birnbaum D, Rosnet O. The centrosomal FOP protein is required for cell cycle progression and survival. Cell Cycle 2009;8:1217–27.
52. Cappell KM, Sinnott R, Taus P, Maxfield K, Scarbrough M, Whitehurst AW. Multiple cancer testis antigens function to support tumor cell mitotic fidelity. Mol Cell Biol 2012;32:4131–40.
53. Izumiyama T, Minoshima S, Yoshida T, Shimizu N. A novel big protein TPRBK possessing 25 units of TPR motif is essential for the progress of mitosis and cytokinesis. Gene 2012;511:202–17.
54. Okada M, Cheeseman IM, Hori T, Okawa K, McLeod IX, Yates JR 3rd, et al. The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. Nat Cell Biol 2006;8: 446–57.
55. Zjablovskaja P, Kardosova M, Danek P, Angelisova P, Benoukraf T, Wurm AA, et al. EVI2B is a C/EBPalpha target gene required for granulocytic differentiation and functionality of hematopoietic progenitors. Cell Death Differ 2017;24: 705–16.
56. Valanejad L, Nadolny C, Shiffka S, Chen Y, You S, Deng R. Differential feedback regulation of Δ4–3-oxosteroid 5β-reductase expression by bile acids. PLoS One 2017;12:e0170960.
57. Traini M, Quinn CM, Sandoval C, Johansson E, Schroder K, Kockx M, et al. Sphingomyelin phosphodiesterase acid-like 3A (SMPDL3A) is a novel nucleotide phosphodiesterase regulated by cholesterol in human macrophages. J Biol Chem 2014;289:32895–913.
58. 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. Sci Rep 2015;5:16286.
59. Quintana-Murci L. Understanding rare and common diseases in the context of human evolution. Genome Biol 2016;17:225.
60. Shields PG, Harris CC. Cancer risk and low-penetrance susceptibility genes in gene-environment interactions. J Clin Oncol 2000;18:2309–15.