


METHODOLOGY ARTICLE

Open Access



# Stepwise approach to SNP-set analysis illustrated with the MetaboChip and colorectal cancer in Japanese Americans of the Multiethnic Cohort

John Cologne<sup>1\*</sup> , Lenora Loo<sup>2</sup>, Yurii B. Shvetsov<sup>2</sup>, Munechika Misumi<sup>1</sup>, Philip Lin<sup>2</sup>, Christopher A. Haiman<sup>3</sup>, Lynne R. Wilkens<sup>4</sup> and Loïc Le Marchand<sup>2</sup>

## Abstract

**Background:** Common variants have explained less than the amount of heritability expected for complex diseases, which has led to interest in less-common variants and more powerful approaches to the analysis of whole-genome scans. Because of low frequency (low statistical power), less-common variants are best analyzed using SNP-set methods such as gene-set or pathway-based analyses. However, there is as yet no clear consensus regarding how to focus in on potential risk variants following set-based analyses. We used a stepwise, telescoping approach to analyze common- and rare-variant data from the Illumina MetaboChip array to assess genomic association with colorectal cancer (CRC) in the Japanese sub-population of the Multiethnic Cohort (676 cases, 7180 controls). We started with pathway analysis of SNPs that are in genes and pathways having known mechanistic roles in colorectal cancer, then focused on genes within the pathways that evidenced association with CRC, and finally assessed individual SNPs within the genes that evidenced association. Pathway SNPs downloaded from the dbSNP database were cross-matched with MetaboChip SNPs and analyzed using the logistic kernel machine regression approach (logistic SNP-set kernel-machine association test, or sequence kernel association test; SKAT) and related methods.

**Results:** The TGF- $\beta$  and WNT pathways were associated with all CRC, and the WNT pathway was associated with colon cancer. Individual genes demonstrating the strongest associations were *TGFBR2* in the TGF- $\beta$  pathway and *SMAD7* (which is involved in both the TGF- $\beta$  and WNT pathways). As partial validation of our approach, a known CRC risk variant in *SMAD7* (in both the TGF- $\beta$  and WNT pathways: rs11874392) was associated with CRC risk in our data. We also detected two novel candidate CRC risk variants (rs13075948 and rs17025857) in *TGFBR2*, a gene known to be associated with CRC risk.

**Conclusions:** A stepwise, telescoping approach identified some potentially novel risk variants associated with colorectal cancer, so it may be a useful method for following up on results of set-based SNP analyses. Further work is required to assess the statistical characteristics of the approach, and additional applications should aid in better clarifying its utility.

**Keywords:** Colorectal cancer, Genome-wide association study, MetaboChip array, Multiethnic Cohort, SNP set analysis

\* Correspondence: [jcologne@ref.jp](mailto:jcologne@ref.jp)

<sup>1</sup>Department of Statistics, Radiation Effects Research Foundation, Hiroshima 732-0815, Japan

Full list of author information is available at the end of the article



## Background

In the search to uncover the missing heritability of complex human diseases [1, 2], agnostic analyses of genome-wide SNP array or sequencing data are giving way to SNP-set or gene-set analyses using methods such as burden tests and kernel association tests. Although such combined-locus approaches are well understood, how to proceed to the next step of identifying the individual risk loci is an area that remains open for development. Our aim was to implement existing set-based testing methods and then use the results to focus in on genes and SNPs to seek new candidate risk variants, without imposing rigid testing criteria, in the spirit of relying on independent external replication as the best way of establishing association [3].

Analyses of individual SNPs, especially less-common ones, suffer from lack of power because multiple-testing adjustment is conservative and individual variants may have low frequencies and small-to-moderate effects. It has therefore been suggested that analyses of groups of variants (both common and rare) that contribute to a common mechanism may be more likely to explain common diseases at the population level, with different variants acting (being present) in different individuals [4]. Across the population, then, many variants might contribute to disease risk via a common pathway or cellular network, so in population-based association analyses power can be increased by combining SNPs within genes or pathways, and treating the SNP set, rather than the individual SNP, as the unit of risk [5–9]. This is especially advantageous with rare variants, because their low frequencies render individual-variant approaches unsuitable [10, 11], although due to low population frequency, aggregate analyses with rare variants might still require greater sample sizes than traditional genome-wide analyses of individual common variants [12]. In addition, an alternative to larger sample sizes and dealing with the challenge of multiple comparisons is to incorporate prior information in the selection of SNP sets for study, to exclude likely non-informative (neutral) loci and thereby increase power by reducing dimensionality and the burden of strict multiple-testing adjustment. We undertook such an approach to study how genetic variants based on genotype data from the MetaboChip [13] are associated with colorectal cancer (CRC) in the individuals of Japanese ancestry included in the Multiethnic Cohort (MEC), a prospective cohort study including five racial/ethnic populations (White, Latino, African-American, Japanese-American, and Native Hawaiian) conducted in Honolulu and Los Angeles [14]. Although the MetaboChip is not a high-density array, it focuses on a number of metabolic pathways—such as those associated with insulin resistance, lipid metabolism, and obesity—that are thought to be involved in CRC etiology.

CRC is an important target for genomic study because it ranks among the top contributors worldwide to cancer incidence and mortality, with substantial differences by ethnic group and involvement of dietary and other lifestyle factors [15–17]. As of 2014, CRC was the third most common cancer and third leading cause of cancer death in both men and women in the U.S. [18], and as of 2012 CRC rates in Miyagi Prefecture, Japan, were the highest among a worldwide selection of registries [16]. Anywhere from 5 to 10% [19] to as much as 15–30% [20] of CRC may be due to known hereditary conditions, including hereditary non-polyposis colorectal cancer (HNPCC; also known as Lynch syndrome, caused by mutations in mismatch DNA repair genes) and familial adenomatous polyposis (FAP, which is caused by mutations in the *APC* tumor suppressor gene). The remainder, sporadic CRC, is commonly attributed to environmental factors, such as a high-caloric, low-fiber, low-calcium western-type diet, low physical activity, obesity, alcohol, and smoking, which presumably involve interactions with predisposing genomic variants [21]. Importantly, offspring of Japanese migrants to Hawaii have had increased rates of CRC far exceeding rates in Japan and even higher than rates in the white population [19, 22]. In recent decades, CRC rates in Japan have increased markedly and have now reached levels that are the same as, or higher than, rates in the United States [23]. Although much of the high incidence of CRC in Japanese is attributed to environmental factors, it is likely that gene-environment interaction also plays a role [24].

The goal of our investigation was to evaluate the use of SNP-set analysis as a preliminary step in ultimately focusing in on potential risk variants. By using what might be called a “telescoping” approach, we began with candidate pathways to limit the initial search for risk variants, then we focused in on genes within the pathways that evidenced association, and finally we zeroed in on variants within the genes that appeared to be associated. Although not a rigorous procedure from the standpoint of statistical testing, such an approach is expected to have greater power to identify potential causal variants than whole-genome testing based on individual-SNP analyses, if it is followed by independent studies focused on the candidate variants.

## Methods

### Study population and genotyping

The MEC, comprising more than 200,000 persons, was assembled in 1993–1996 by the mailing of a self-administered, 26-page questionnaire to persons with drivers licenses (California and Hawaii), voter registrations (Hawaii only), or health care financing records (California only) to obtain extensive information on demographics, medical and reproductive histories, medication use, family history of various cancers, physical activity, and diet. Ancestry in the MEC was ascertained via questionnaire [25].

Because the importance of certain cellular pathways might vary due to ethnic differences, focusing on persons of a single ancestry should be advantageous by reducing variability. We therefore restricted our analysis to persons of Japanese ancestry, for reasons explained in the Background section. The Japanese-American sub-population constitutes about 26% of the MEC.

Identification of incident cancer cases was by regular linkage with the Hawaii, Los Angeles County, and California SEER registries. Although colon and rectal cancers are distinct and have separate ICD codes, they are often combined because their etiologies are similar. In the present analysis we used all CRC and colon cancer only; we did not analyze rectal cancer alone due to the small number of cases.

Genotyping was performed in blood specimens collected according to a case-control design. Some MEC subjects were re-contacted, mostly from 1995 to 2001, for blood collection; these included persons with incident breast, prostate, or colorectal cancers, as well as a random sample of cohort participants to serve as controls in nested genetic case-control studies (participation rate 72% among cases and 63% among controls). From 2001 to 2006, blood was also collected prospectively, without regard to cancer diagnosis, from willing cohort participants (participation rate 43%).

Genotypes were assessed with the MetaboChip, a custom Illumina iSelect array designed with about 200,000 SNPs to study genetic association with metabolic, cardiovascular, and anthropometric traits. The MetaboChip was not designed to study cancer, but it includes variants known or suspected to be associated with metabolism, obesity, and insulin resistance—factors that have been linked to CRC risk. Although the MetaboChip has limited coverage of the genome, larger agnostic GWAS arrays are likely to include large numbers of non-informative SNPs, which can reduce power in gene-set analyses [26]. In addition, although imputation allows estimation of many non-genotyped variants, imputation is challenging with rare variants [26] and imputation with persons of Japanese ancestry in the MEC is based on East Asians in the 1000 Genomes Project [27], which might not be the most suitable basis for imputation of rare variants among persons of strictly Japanese ancestry. We therefore considered that the MetaboChip genotype data (without imputation) could be useful for a preliminary examination of CRC pathways because of its focused nature and direct genotyping of less-common variants.

The study protocols of the MEC GWASs were approved by the University of Hawaii Human Studies Program and the University of Southern California IRB.

#### Data pre-processing

We chose candidate pathways for the first step (pathway analysis) by assessing published reviews of molecular characteristics of CRC and selecting pathways and their related

genes that were anticipated to be associated with CRC, a strategy that is expected to improve power [28, 29]. A similar approach was also described by Liu and others [30]. Molecular characteristics of CRC are the subject of several reviews [20, 31–33]. Pathways chosen were WNT, TGF-beta, P53, RTK-RAS, MAPK, adiponectin, combined DNA repair and fidelity of DNA replication, mTOR, and the laminin gene family. Genes we selected from among these pathways are listed in the Additional file 1. We downloaded lists of SNPs in the selected genes from the dbSNP database [34], similar to the approach of Scarborough and others [8], except that we did not restrict upstream and downstream distances of candidate variants (it is not clear how association tests with aggregated variants will perform with non-coding variants [12] possibly involved in regulation, so to err on the side of not leaving anything out, we chose to include as many variants as possible). Lists of SNP rs numbers in the selected genes were queried with restriction to “Organism: *Homo sapiens*” and “Variation class: SNP” in the dbSNP database. We copied the list of rs numbers shown in the “dbSNP Batch” option of “Display Settings”. The “dbSNP Batch” list includes rs numbers of SNPs that have been merged with other SNPs, which is important given the time that has elapsed between specification of variants for the MetaboChip array and downloading of SNP lists (older, merged rs numbers were not included in the “FlatFile” option of dbSNP).

The downloaded lists of rs numbers were then matched against the rs numbers of variants in the MetaboChip data to create the list of variants for analysis (an R script for this processing is available upon request). Downloads were current as of July 28, 2016 or later and were based on genome build 38. Matching SNPs were identified in all pathways except for the laminin gene family. Numbers of SNPs that matched to the MetaboChip are shown in the Additional file 1.

We excluded cohort participants whose reported sex did not match the sex chromosome genotype, whose overall genotype call rate was less than 95%, who were first-degree relatives, or whose genotype was found to be duplicated. Of the 8187 remaining participants of Japanese ancestry, 331 were ineligible due to prior cancer, leaving 7856 subjects for analysis: 676 with colorectal cancer (478 with colon cancer) and 7180 controls. We included all genotyped loci from the MetaboChip that remained after we excluded markers not in Hardy-Weinberg equilibrium, markers with call rate less than 95%, and sex-chromosome and mitochondrial DNA SNPs. After these variant exclusions there were 189,127 MetaboChip SNPs available for matching to the downloaded pathway SNPs. Variants with minor allele frequency (MAF) less than 1% were retained for the present analyses because genes containing common variants with established effects on complex diseases might also contain rare variants with larger effects [1].

### Analysis

To implement the telescoping approach, in the first step we applied the Sequence Kernel Association Test (SKAT [35], formerly known as the logistic SNP-set kernel-machine association test [29]) to pathways as units of analysis. In the second step we applied SKAT to the set of genes within each pathway (one pathway at a time) that showed evidence of association with CRC. In the third step, individual SNPs contained within the genes that evidenced association in analyses with SKAT were analyzed with PLINK [36]. Logistic models (defined below) were fit in SKAT and PLINK with the one-parameter linear genetic effect (count of minor-variant alleles: 0, 1, or 2) as the genomic covariate. *P* values corrected for multiple testing were obtained with the Bonferroni family-wise error rate (FWER) and the false discovery rate (FDR) procedures (note that FDR is not necessarily preferable to FWER in situations with a small number of tests, such as when confirming results in an independent candidate-SNP study [37]). Individual SNPs that showed evidence of association were further examined by searching for their rs numbers in the NHGRI-EBI GWAS Catalogue [38] and in PubMed.

Pathway analyses were performed with the kernel logistic regression procedure in the SKAT R package (v. 1.2.1) [39], which can accommodate rare variants [35]. Briefly, SKAT is based on a variance component score statistic, where the variance component is the variance—within a pathway—of individual-variant effects assumed to follow a common distribution, so that the null hypothesis—that all individual-variant effects are zero—is equivalent to the simpler hypothesis that the variance of those effects is zero. For a binary phenotype (outcome)  $Y \in \{0,1\}$ , the logistic regression model is

$$\text{logit Pr}(Y_i = 1) = \alpha' x_i + \beta' g_i, \tag{1}$$

where logit is the logistic function  $\{\text{logit}(p) = \log(p/[1-p])\}$ , the subscript *i* signifies individual *i* in the study population ( $i = 1, \dots, n$ ),  $x_i = \{1, x_{i1}, \dots, x_{ip}\}'$  is the vector of covariates for individual *i*,  $g_i = \{g_{i1}, \dots, g_{iq}\}'$  is the vector of genotypes at *q* SNPs for individual *i*,  $\alpha$  is a  $(p + 1) \times 1$  vector of coefficients for the covariates, and  $\beta$  is a  $q \times 1$  vector of coefficients for the SNPs. The variance component statistic (*S*) is

$$S = (\mathbf{y} - \hat{\mathbf{y}}) \mathbf{G} \mathbf{W} \mathbf{G}' (\mathbf{y} - \hat{\mathbf{y}}),$$

where  $\mathbf{y} = \{y_1, \dots, y_n\}'$  is the vector of observed outcomes,  $\hat{\mathbf{y}}$  is the vector of fitted values under the null hypothesis ( $\hat{\mathbf{y}} = \text{logit}^{-1}[\mathbf{X}\hat{\alpha}]$ ,  $\mathbf{X} = [x_1, \dots, x_n]'$ ),  $\mathbf{G} = [g_1, \dots, g_n]'$ , and  $\mathbf{W}$  is a  $q \times q$  diagonal matrix of individual-variant weights that can be chosen to improve power (e.g., by down-weighting non-functional variants [35]). We used

an approach similar to that used by Saunders and others [40], in that we ran alternative analyses with various omnibus tests, such as the optimized combination of burden and SKAT tests (SKAT-O) [41]—which has better power than traditional burden tests [42]—and a test for combined rare and common variants (SKAT-C) [43]. We used the default linear-weighted method in SKAT, which assigns higher weights to rarer variants due to their greater likelihood of being causal.

Covariates we adjusted in logistic regression models were age (*a*), sex (*s*), body mass index (BMI, the Quetelet index  $q = \text{height}/\text{weight}^2$ ), smoking behavior (*c*), and, to account for population stratification, the top five ancestry-informative eigenvectors ( $p_1, \dots, p_5$ ) from the principal component decomposition of the genotype matrix among MEC Japanese. Logistic regression models for a single variant ( $j \in \{1, \dots, q\}$ ) were therefore of the form

$$\begin{aligned} \text{logit}[\text{Pr}_i(\text{cancer})|x_i, g_{ij}] = & \alpha_0 + \alpha_a a_i + \alpha_s s_i + \alpha_q q_i \\ & + \alpha_h h_i + \alpha_c c_i + \sum_{k=1}^5 \alpha_{p_k} p_{ik} \\ & + \beta_g g_{ij} \end{aligned}$$

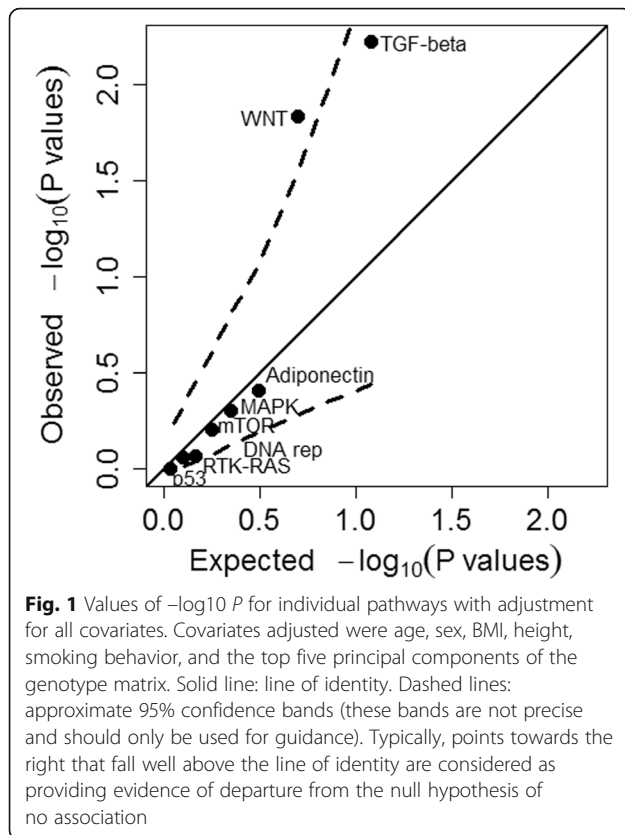
where  $\text{Pr}_i(\text{cancer})$  is the probability that individual *i* has colorectal cancer (or colon cancer, depending on which outcome is being analyzed) and  $g_{ij}$  is the *j*th genomic covariate (count of minor alleles) for individual *i*. Logistic models used in SKAT had the obvious multi-locus extension of the genomic covariate, as shown in eq. (1). Height (*h*) was included to remove possible residual dependence of BMI on stature [44], although it had little impact on the results.

SKAT produces a *P* value plot (QQ plot based on the expected uniform distribution of *P* values under the global null hypothesis) as a means of assessing true positives [45]. The *P* values in SKAT are adjusted for poor adherence to asymptotic test assumptions with a binary outcome but are not a priori adjusted for multiple testing (multiple SNP sets).

### Results

#### Pathway and individual-gene results

With adjustment for all covariates, the TGF- $\beta$  ( $P = 0.0060$ ) and WNT ( $P = 0.015$ ) pathways demonstrated associations with all CRC (Fig. 1). With multiple testing adjustment by the Bonferroni and FDR procedures, only the TGF- $\beta$  pathway evidenced association (Bonferroni  $P = 0.048$  based on eight tests, Benjamini & Hochberg FDR  $P = 0.048$  with the R p.adjust [“method = fdr”] procedure; corresponding values for the WNT pathway were 0.12 and 0.058).



With the bootstrap facility in SKAT to adjust for multiple testing, the family-wise error rate adjusted  $P$  value for TGF- $\beta$  was 0.087 and that for WNT was 0.17. The SKAT-C method produced unadjusted  $P$  values 0.052 for WNT and 0.90 for TGF- $\beta$  (the TGF- $\beta$  and WNT pathways had similar proportions of common and rare SNPs; see Appendix 3). The SKAT-O combination of traditional-burden and SKAT methods produced  $P$  values 0.012 for TGF- $\beta$  and 0.018 for WNT. Given that the TGF- $\beta$  and WNT pathways are closely related [46], we also performed a pathway analysis with TGF- $\beta$  and WNT combined as a single pathway; the combined pathway had an unadjusted  $P = 0.0023$  (corrected  $P = 0.016$  by both Bonferroni and FDR).

Several genes demonstrated evidence of association with CRC when SKAT was applied to genes as SNP sets within either the TGF- $\beta$  or the WNT pathway (Table 1). None of these genes evidenced association after Bonferroni correction: significance thresholds were  $0.05/12 = 0.0042$  for TGF- $\beta$  genes and  $0.05/13 = 0.0038$  for WNT genes (numbers of genes—12 and 13—that overlapped with the MetaboChip were derived from Additional file 1: Table S1). The smallest  $P$  value was for *WNT11* (based on only one overlapping SNP), whereas the gene with the next smallest  $P$  value (*TGFBR2*) contained 78 overlapping SNPs.

**Table 1** Individual genes in the TGF- $\beta$  and WNT pathways that demonstrated an association with colorectal cancer

Gene	No. SNPs	Pathway	$P$ value	
			All CRC	Colon cancer only
<i>WNT11</i>	1	WNT	0.0064	0.014
<i>TGFBR2</i>	78	TGF- $\beta$	0.0075	0.15
<i>SMAD7</i>	6	TGF- $\beta$ and WNT	0.015	0.058
<i>TCF7L2</i>	26	WNT	0.019	0.045
<i>TFDP1</i>	1	TGF- $\beta$	0.044	0.040

In an analysis with colon cancer only as the endpoint, the WNT pathway appeared to be associated ( $P = 0.045$  not corrected for multiple testing) but the TGF- $\beta$  pathway did not ( $P = 0.18$ ). With the SKAT-C method,  $P = 0.11$  for the WNT pathway, and with the SKAT-O method  $P = 0.036$  for the WNT pathway. Within the WNT pathway, individual genes evidencing association with colon cancer were *WNT11* ( $P = 0.014$ ) and *TCF7L2* ( $P = 0.045$ ). *SMAD7* showed only weak evidence of association with colon cancer ( $P = 0.058$ ). The *TFDP1* gene in the TGF- $\beta$  pathway demonstrated an association with colon cancer even though the TGF- $\beta$  pathway overall did not. None of these putative associations with colon cancer would be deemed statistically significant after correction for multiple testing, however. For this reason, and because pathway results were qualitatively similar for all CRC and colon cancer only, we used only CRC in further analyses of individual variants.

#### Individual-SNP results

Eleven individual MetaboChip SNPs in the TGF- $\beta$  and WNT pathways were associated with CRC in unadjusted case-control association analysis (Table 2; two SNPs—rs11874392 and rs4464148—are in both pathways). With covariate adjustment, three SNPs in the TGF- $\beta$  pathway (rs17025857 and rs13075948 in the *TGFBR2* gene, and rs11874392 in the *SMAD7* gene) were associated with CRC; one of these (rs11874392) is also in the WNT pathway, but no other SNPs in the WNT pathway were associated with CRC after covariate adjustment. With restriction to the 405 MetaboChip SNPs in our chosen pathways, the minimum value of FDR was 0.41 (Table 2); Bonferroni adjustment resulted in significance levels of 1.0 for all 405 SNPs.

The variant rs11874392 in *SMAD7* (adjusted OR 1.16) was reported by Jiang and others [47] to be associated with CRC in a population-based study of non-Hispanic white subjects, but with odds ratio less than 1 (OR 0.80), whereas it was noted to be positively associated with CRC in Hispanics by Schmit and others (OR = 1.27) [48]. The variant rs13075948 in the *TGFBR2* gene (OR 1.95) is an intron variant that has been implicated in abdominal aortic aneurism [49], but it did not otherwise return any results in searches on NHGRI and PubMed,

**Table 2** Association with colorectal cancer of individual TGF- $\beta$  and WNT pathway SNPs

SNP	Gene	Minor allele frequency		Crude (unadjusted) association				Adjusted association <sup>a</sup>				
		Cases	Controls	Odds ratio	95% CI	P value	FDR <sup>c</sup>	Odds ratio	95% CI	P value	FDR	
<i>TGF-<math>\beta</math> pathway</i>												
rs17025857	<i>TGFBR2</i>	0.029	0.019	1.57	1.12, 2.21	0.009	0.48	2.01	1.28, 3.16	0.002	0.41	
rs13075948	<i>TGFBR2</i>	0.026	0.018	1.49	1.04, 2.13	0.029	0.58	1.95	1.22, 3.11	0.005	0.44	
rs11874392	<i>SMAD7</i>	0.375	0.342	1.16	1.03, 1.30	0.015	0.49	1.16	1.00, 1.34	0.049	1	
rs3825977	<i>SMAD3</i>	0.453	0.490	0.86	0.77, 0.96	0.009	0.48	0.87	0.76, 1.01	0.063	1	
rs4776890	<i>SMAD3</i>	0.206	0.234	0.85	0.74, 0.97	0.019	0.49	0.85	0.72, 1.01	0.059	1	
rs4464148	<i>SMAD7</i>	0.038	0.053	0.70	0.53, 0.94	0.016	0.49	0.71	0.50, 1.01	0.057	1	
rs11466531	<i>TGFBR2</i>	0.011	0.019	0.57	0.34, 0.97	0.034	0.59	0.71	0.39, 1.28	0.26	1	
rs3773662	<i>TGFBR2</i>	0.011	0.020	0.56	0.33, 0.94	0.026	0.56	0.68	0.38, 1.23	0.20	1	
<i>WNT pathway</i>												
rs2439593	<i>APC</i>	0.0022	0.0003	6.39	1.53, 26.8	0.004	0.48	NA <sup>b</sup>	NA	NA	NA	
rs4944092	<i>WNT11</i>	0.156	0.129	1.24	1.07, 1.45	0.006	0.48	1.13	0.93, 1.38	0.22	1	
rs11874392	<i>SMAD7</i>	0.375	0.342	1.16	1.03, 1.30	0.015	0.49	1.16	1.00, 1.34	0.049	1	
rs4464148	<i>SMAD7</i>	0.038	0.053	0.70	0.53, 0.94	0.016	0.49	0.71	0.50, 1.01	0.057	1	
rs11196187	<i>TCF7L2</i>	0.023	0.033	0.68	0.47, 0.99	0.042	0.60	0.69	0.43, 1.12	0.13	1	

<sup>a</sup>Adjusted for age, sex, BMI, height, smoking status, and the first five principal components among Japanese; odds ratios based on a linear genetic effect

<sup>b</sup>NA: could not be estimated because of low frequency in controls

<sup>c</sup>FDR: false discovery rate based on the 405 MetaboChip SNPs that were in the selected pathways

nor did the variant rs17025857, also an intron variant in the *TGFBR2* gene (adjusted OR 2.01), although other variants in the *TGFBR2* gene have been reported to be associated with CRC [50, 51].

We compared frequencies of the SNPs we detected as being associated with CRC in our population of Japanese ancestry with those of the Tokyo Japanese population (JPT; 120 samples) in the 1000 Genomes Catalog (using the NCBI 1000 Genomes Browser, Phase 3 [52]). All three minor allele frequencies in our population were slightly higher than those in the Tokyo Japanese: SNP rs17025857 (our cohort MAF 0.020) had MAF 0/120 in the Tokyo Japanese; rs13075948 (our cohort MAF 0.019) had MAF 0/120 in the Tokyo Japanese; and rs11874392 (our cohort MAF 0.34) had MAF 0.26 in the Tokyo Japanese.

## Discussion

Agnostic (individual-variant) approaches to association testing can suffer from a lack of statistical power due to low variant frequency and moderate-at-best effect sizes. SNP-set or gene-set analyses (pathway analyses) based on burden tests or kernel association tests are a more powerful approach but do not reveal individual causal SNPs. Recent methodological work has turned to this problem of identifying causal variants after set-based testing, but the area remains open for further development. Based on the premise that the most valid approach to confirming relationships between variants and disease is to conduct independent external replication, a variant-discovery approach that identifies candidates for

further, independent investigation should be a useful first stage in identifying risk variants (see Robertson and others [53] for recent work on combining data from two-stage studies). In the present work we illustrate a multi-step, telescoping approach that is motivated not by rigorous significance testing but rather by sequentially removing natural layers of complexity in the analysis. We suggest some variants that might deserve further study in relation to colorectal cancer, but our illustration is not meant to be conclusive with regard to the association of these variants with CRC. The approach can be applied to genome-wide array data or whole-exome (or whole-genome) sequencing data with the intention to follow up results with independent data (including in silico studies).

Our analysis of SNP sets with the SKAT method, which began by limiting the set of candidate SNPs to those in pathways having a known mechanistic role in CRC, identified several variants in the TGF- $\beta$  and WNT signaling pathways that are potentially associated with CRC in Japanese Americans. Two of those variants (rs17025857, OR = 2.01, and rs13075948, OR 1.95, both in the *TGFBR2* gene) are apparently new findings given that their association with CRC was not noted in the NHGRI-EBI GWAS Catalog or in PubMed. TGF- $\beta$  and WNT pathway proteins influence cell division and cell fate of gut endoderm stem cells, such that disorders in these pathways can lead to gastro-intestinal cancers, including colonic adenocarcinomas [46]. Association tests using gene sets within these pathways confirmed that *TGFBR2* in the TGF- $\beta$  pathway,

and *SMAD7* in both pathways, are associated with CRC. Although no individual variants in our analysis would be considered statistically significant based on traditional multiple-testing adjustment methods, the purpose of our analysis was to find previously unidentified candidate risk variants within pathways and genes already known to be related to CRC, so strict *P*-value adjustment—which is conservative and may result in false-negative results (type II errors)—might not be appropriate. Correcting the family-wise error rate assumes a global null hypothesis of no associated elements, whereas it is likely that some, and perhaps many, of the SNPs in the chosen pathways are associated with CRC. By using a stepwise approach, starting with candidate pathways and then telescoping in on genes and then SNPs within genes that demonstrate evidence of association with CRC, some of the overly conservative restrictiveness of traditional multiple testing (and resulting low probability of detecting risk variants) may be overcome. However, such a multi-step approach could still be subject to inflated type I errors, so our findings should only be considered preliminary. As with all genomic analyses, the most important evidence must come from independent confirmation in independent populations. Low density of coverage of the genome might limit the effectiveness of pathway analysis [54], which may be a reason why our selected pathways other than TGF- $\beta$  and WNT, also known to be associated with CRC, did not show evidence of association in our analysis. Indeed, among our selected pathways, the TGF- $\beta$  pathway had the largest number of SNPs present on the MetaboChip, so it might not be surprising that it produced the strongest evidence of association.

There has been little guidance on how to identify individual driver (risk) SNPs that underlie a SNP set found to be related to phenotype. Various ad hoc approaches have been used; for example, Tang and others [55] chose SNPs that were deemed to be associated by individual-SNP analysis and in genes that were deemed to be associated via SKAT-O analysis. Recently, He and others [9] described a variable-selection method incorporated within the SKAT kernel approach that can be used to suggest which SNPs drive the SNP-set association. In particular, finding a SNP set that evidences association does not distinguish between a few driver SNPs with large effects on the one hand and many driver SNPs with lesser effects on the other.

SNP-set analyses present a number of challenges when rare variants are studied. If rare variants in a particular genetic region are enriched among persons with disease, set-based tests should be more powerful than individual-SNP analyses. However, single variants might not contribute greatly to more powerful SNP-set tests if the number of associated variants in any particular gene-set or SNP-set is small. Furthermore, the optimal approach to testing association with SNP sets depends on many factors, including the true (but unknown) proportion of causal

SNPs in the SNP set and their effect magnitudes. We therefore employed several approaches (SKAT, SKAT-C, and SKAT-O), but which is most appropriate cannot be known a priori. Another concern is that, with small samples, the asymptotic distribution used for the SKAT test might be inaccurate. However, this has been said to result in inflated *P* values (or loss of power [56]), in which case small-sample bias would not likely cause false-positive results. It is also possible that variant assignment to a particular gene might be incorrect [57].

Appropriate weighting of individual SNPs within SNP sets would be beneficial if it were feasible. In addition to using biological (functional) information for pathway and gene selection, using such information to weight individual variants in the analysis could reduce the impact of non-informative variants within the selected genes and pathways. We used the default weighting in SKAT, which gives higher weight to rarer variants, but there is surely much residual variation in strength of effect even after accounting for variant frequency. More appropriate weights that take function into account might lead to increased power by down-weighting neutral variants that have little or no impact on cellular processes. However, such functional and annotation information may be advantageous only when it is accurate [2], which remains a problem with online databases. Spencer and others [58] described an alternative, Bayesian, approach that allows incorporation of expert prior functional knowledge.

Several strengths and limitations of the present study deserve consideration. The fact that we were able to reproduce associations with previously reported CRC-related genes provides reassurance about the validity of our approach. A second strength is the design of the MEC study, a prospective cohort study with population based sampling. A third strength is that we focused on one ethnic group: differences in allelic variation among populations can result in reduced power if association tests are not performed in specific populations [59]. A fourth strength is that, by focusing on pathways and genes known to be mechanistically linked to CRC, it is more likely that a true risk variant might be detected, because the genes considered should be enriched with variants related to CRC, and with lower degrees of freedom there is less penalty for multiple testing and hence a lower false-negative rate. One limitation of this study is the relatively small number of cases, which could reduce power. A second limitation is sparse coverage of the genome by the MetaboChip for many of the pathways and genes considered in our analysis. There may well be variants in these genes that are associated with CRC but that were not included in the present study because they were not genotyped. A more comprehensive SNP set including imputed variants could be more informative, but as noted in the introduction, imputation with rare variants, especially among persons of Japanese

ancestry, may be unreliable. Alternatively, SNP-set analyses based on whole-genome sequence data can be employed to include and potentially discover novel CRC-related genes, as was done by Koboldt and others [60], who combined whole-genome sequencing with selection of known susceptibility genes for prostate cancer. A third limitation is the small set of candidate pathways selected. For example, we did not examine DNA mismatch repair gene defects, which have been linked to HNPCC but are rare in sporadic CRC [20], genes related to the inflammation and innate immunity pathways [61], or genes related to glucose metabolism and its interaction with epithelial-mesenchymal transition [62]. Furthermore, we did not investigate variants in genes or pathways not yet known to be associated with CRC. Better informed decisions as to which pathways to include in the analysis—rather than limiting to pathways with demonstrated associated SNPs—could perhaps further increase the likelihood of detecting novel risk SNPs, because there could be many false negative results among published GWAS studies [63]. However, adding candidate pathways could also increase the proportion of non-informative (neutral) SNPs. A fourth limitation is the lack of a well-defined statistical-testing framework for the telescoping approach we used. Larson and others [64] pointed out that type-I error control has not been well studied for gene-set analyses, and they described corrections to SKAT to adjust for multiple testing when there may be substantial overlap of genes across multiple pathways. However, because of our small number of candidate pathways, there was only modest overlap of genes. Although our approach might suffer from lack of tight type-I error control, it is less likely to reject true causal variants, so it should be useful as a first step in identifying candidate SNPs to be assessed in a second stage of independent validation.

Although we noted potential associations of several variants with CRC, one should keep in mind that a mechanism comprises the collective effects of numerous individual minor variants across the population. It is not finding the individual risk variant per se that is the ultimate goal. Rather, the variants identified should be considered as proxies for the entire mechanism in which they participate, and other variants that impact that mechanism should be considered likely candidates (with effects—and therefore strength of associations—dependent, of course, on their relative functionalities). Indeed, it has been noted that with the explosion of the human population, there are likely to have arisen many rare variants that might play roles in complex disease risk, but with only a few individuals in any particular study sample possessing any one particular variant among them [65]. In this regard, the estimated odds ratios for individual variants may be more informative than their *P* values for association.

## Conclusions

A stepwise, telescoping approach to the analysis of dense genomic data—beginning with pathways, then focusing in on genes within the pathways associated with outcome, and finally assessing individual SNPs within the genes that evidence association with outcome—allowed us to identify several potential novel risk variants not previously associated with CRC in traditional analyses. The procedure is exploratory, so these variants require independent validation as well as consideration of their cellular or regulatory functions before they can be regarded as causal SNPs. Our results are meant merely to demonstrate the potential utility of the stepwise approach to SNP-set analyses; the procedure should identify a greater number of potential associated variants if based on a genome-wide scan. Although multiple-testing implications of the stepwise approach could be complex, this type of approach—coupled with subsequent independent confirmation as well as detailed functional considerations—may be preferable to the agnostic approach typically employed with whole-genome scans. Furthermore, additional novel candidate SNPs might be identified if the initial set of candidate pathways is expanded to include ones that are hypothesized to be associated on the basis of functional or biological knowledge even though they have not previously been established as being associated with CRC (due to low power of agnostic analyses). Future, more in-depth, analyses of CRC risk pathways, genes, and variants should be based on denser coverage of the genome, include a larger number of candidate pathways, investigate differences across ethnic populations, and utilize imputed variants that are imputed with reasonable certainty. Because a large number of neutral loci can dilute statistical power, a potentially fruitful area of future research would be to incorporate functional information to weight pathways, genes, and individual variants according to biological expectations of their relevance to the outcome under study.

## Additional file

**Additional file 1:** Detailed characterization of genes and pathways used in the analysis. Supplementary Material. Detailed characterization of genes and SNPs used in the analysis. (DOCX 51 kb)

## Abbreviations

BMI: body mass index; CRC: colorectal cancer; MAF: minor allele frequency; MEC: Multiethnic Cohort; SKAT: Sequence Kernel Association Test; SNP: single nucleotide polymorphism

## Acknowledgements

The first author is grateful to the faculty and staff of the University of Hawaii Cancer Center (UHCC) for their hospitality and guidance during a brief sabbatical, at which time the present work was initiated.



### Funding

This work was supported in part by US National Institute of Health grants U01 CA164973 and U01 HG004802. The Radiation Effects Research Foundation (REFR), Hiroshima and Nagasaki, Japan is a public interest incorporated foundation funded by the Japanese Ministry of Health, Labour and Welfare (MHLW) and the US Department of Energy (DOE). This publication was supported in part by REFR Research Protocol 4-04. The views of the authors do not necessarily reflect those of the two governments.

### Availability of data and materials

The data that support the findings of this study are available from the University of Hawaii Cancer Center, but restrictions apply to the availability of these data, which were used under a Material Use Agreement executed between the University of Hawaii and the Radiation Effects Research Foundation for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the investigators responsible for the Multiethnic Cohort Study.

### Authors' contributions

JC merged the SNP lists with the MetaboChip data and performed the analyses. LL selected the relevant pathways. YS prepared the genomic and case-control data for analysis, including writing PLINK scripts for data cleaning. MM provided computational support for the analyses. PL performed the SNP list downloads. CH, LW, and LLM supervised the study. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

All samples were collected with written, signed informed consent. All procedures and study protocols were approved by the Institutional Review Boards (IRBs) at the University of Hawaii and the University of Southern California.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Department of Statistics, Radiation Effects Research Foundation, Hiroshima 732-0815, Japan. <sup>2</sup>Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA. <sup>3</sup>Department of Preventive Medicine and Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA. <sup>4</sup>Biostatistics and Informatics Shared Resource, University of Hawaii Cancer Center, Honolulu, HI 96813, USA.

Received: 16 November 2017 Accepted: 29 June 2018

Published online: 09 July 2018

### References

- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747–53. <https://doi.org/10.1038/nature08494>.
- Mechanic LE, Chen H-S, Amos CI, Chatterjee N, Cox NJ, Divi RL, et al. Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. *Genet Epidemiol*. 2012;36(1):22–35. <https://doi.org/10.1002/gepi.20652>.
- Ioannidis JPA, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet*. 2009;10(5):318–29. <https://doi.org/10.1038/nrg2544>.
- Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*. 2009;19(3):212–9. <https://doi.org/10.1016/j.cog.2009.04.010>.
- Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, et al. Pathway analysis by adaptive combination of P-values. *Genet Epidemiol*. 2009;33(8):700–9. <https://doi.org/10.1002/gepi.20422>.
- Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010;11(12):843–54. <https://doi.org/10.1038/nrg2884>.
- Danaher P, Paul D, Wang P. Covariance-based analyses of biological pathways. *Biometrika*. 2015;102(3):533–44. <https://doi.org/10.1093/biomet/asv013>.
- Scarborough PM, Weber RP, Iversen ES, Brhane Y, Amos CI, Kraft P, et al. A cross-cancer genetic association analysis of the DNA repair and DNA damage signaling pathways for lung, ovary, prostate, breast, and colorectal cancer. *Cancer Epidemiol Biomark Prev*. 2015;25(1):193–200. <https://doi.org/10.1158/1055-9965.EPI-15-0649>.
- He Q, Cai T, Liu Y, Zhao N, Harmon QE, Almlí LM, et al. Prioritizing individual genetic variants after kernel machine testing using variable selection. *Genet Epidemiol*. 2016;40(8):722–31. <https://doi.org/10.1002/gepi.21993>.
- Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Pooled association tests for rare variants in exon-sequencing studies. *Am J Hum Genet*. 2010;86(6):832–8. <https://doi.org/10.1016/j.ajhg.2010.04.005>.
- Mao X, Li Y, Liu Y, Lange L, Li M. Testing genetic association with rare variants in admixed populations. *Genet Epidemiol*. 2013;37(1):38–47. <https://doi.org/10.1002/gepi.21687>.
- Auer PL, Reiner AP, Wang G, Kang HM, Abecasis GR, Altshuler D, et al. Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI exome sequencing project. *Am J Hum Genet*. 2016;99(4):791–801. <https://doi.org/10.1016/j.ajhg.2016.08.012>.
- Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The MetaboChip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet*. 2012;8(8):e1002793. <https://doi.org/10.1371/journal.pgen.1002793>.
- Kolonel LN, Henderson BE, Hankin JH, Nomura AMY, Wilkens LR, Pike MC, et al. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol*. 2000;151(4):346–57.
- Boyle P, Leon ME. Epidemiology of colorectal cancer. *Br Med Bull*. 2002;64:1–25.
- Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends – an update. *Cancer Epidemiol Biomark Prev*. 2016;25(1):16–27. <https://doi.org/10.1158/1055-9965.EPI-15-0578>.
- Song M, Hu FB, Spiegelman D, Chan AT, Wu K, Ogino S, et al. Long-term status and change of body fat distribution, and risk of colorectal cancer: a prospective cohort study. *Int J Epidemiol*. 2016;45(3):871–83. <https://doi.org/10.1093/ije/dyv177>.
- Siegel R, DeSantis C, Jemal A. Colorectal cancer statistics, 2014. *CA Cancer J Clin*. 2014;64(2):104–17. <https://doi.org/10.3322/caac.21220>.
- Hagggar FA, Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin Colon Rectal Surg*. 2009;22(4):191–7. <https://doi.org/10.1055/s-0029-1242458>.
- Fearon ER. Molecular genetics of colorectal cancer. *Annu Rev Pathol*. 2011;6:479–507. <https://doi.org/10.1146/annurev-pathol-011110-130235>.
- Slattery ML, Lundgreen A, Herrick JS, Caan BJ, Potter JD, Wolff RK. Diet and colorectal cancer: analysis of a candidate pathway using SNPs, haplotypes, and multi-gene assessment. *Nutr Cancer*. 2011;63(8):1226–34. <https://doi.org/10.1080/01635581.2011.607545>.
- Boyle P, Langman JS. ABC of colorectal cancer: epidemiology. *BMJ*. 2000;321(7264):805–8.
- International Agency for Research on Cancer. GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. Available from: [http://globocan.iarc.fr/Pages/fact\\_sheets\\_cancer.aspx](http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx). Accessed 13 Nov 2017.
- Wang H, Iwasaki M, Haiman CA, Kono S, Wilkens LR, Keku TO, et al. Interaction between red meat intake and NAT2 genotype in increasing the risk of colorectal cancer in Japanese and African Americans. *PLoS One*. 2015;10(12):e0144955. <https://doi.org/10.1371/journal.pone.0144955>.
- Wang H, Haiman CA, Kolonel LN, Henderson BE, Wilkens LR, Le Marchand L, et al. Self-reported ethnicity, genetic structure and the impact of population stratification in a multiethnic study. *Hum Genet*. 2010;128(2):165–77. <https://doi.org/10.1007/s00439-010-0841-4>.
- Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014;95(1):5–23. <https://doi.org/10.1016/j.ajhg.2014.06.009>.
- Wang H, Burnett T, Kono S, Haiman CA, Iwasaki M, Wilkens LR, et al. Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in V771A. *Nat Commun*. 2014;5:4613. <https://doi.org/10.1038/ncomms5613>.
- Huber W, Hahne F. Annotation and metadata. In: Hahne F, Huber W, Gentleman R, Falcon S, editors. *Bioconductor case studies*. New York: Springer; 2008. Chapter 8.
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet*. 2010;86(6):929–42. <https://doi.org/10.1016/j.ajhg.2010.05.002>.

30. C-y L, Wu MC, Chen F, Ter-Minassian M, Asomaning K, Zhai R, et al. A large-scale genetic association study of esophageal adenocarcinoma risk. *Carcinogenesis*. 2010;31(7):1259–63. <https://doi.org/10.1093/carcin/bgq092>.
31. Network TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330–7. <https://doi.org/10.1038/nature11252>.
32. Colussi D, Brandi G, Bazzoli F, Ricciardiello L. Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention. *Int J Mol Sci*. 2013;14(8):16365–85. <https://doi.org/10.3390/ijms140816365>.
33. Peters U, Bien S, Zubair N. Genetic architecture of colorectal cancer. *Gut*. 2015;64(10):1623–36. <https://doi.org/10.1136/gutjnl-2013-306705>.
34. The National Center for Biotechnology Information. dbSNP Short Genetic Variations database. <https://www.ncbi.nlm.nih.gov/SNP/>. Accessed 13 Nov 2017.
35. Wu MC, Lee S, Cai T, Lee Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82–93. <https://doi.org/10.1016/j.ajhg.2011.05.029>.
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–75. <https://doi.org/10.1086/519795>.
37. Brinster R, Köttingen A, Tayo BO, Schumacher M, Sekula P, CKDGen Consortium. Control procedures and estimators of the false discovery rate and their application in low-dimensional settings: an empirical investigation. *BMC Bioinformatics*. 2018;19(1):78. <https://doi.org/10.1186/s12859-018-2081-x>.
38. The National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI). GWAS Catalog: The NHGRI-EBI Catalogue of published genome-wide association studies. <http://www.ebi.ac.uk/gwas/>. Accessed 14 Nov 2017.
39. Lee S. SKAT Package. 2017. Available from: <https://cran.r-project.org/web/packages/SKAT/vignettes/SKAT.pdf>. Accessed 13 Nov 2017.
40. Saunders EJ, Dadaev T, Leongamornlert DA, Al Olama AA, Benlloch S, Giles GG, et al. Gene and pathway level analyses of germline DNA-repair gene variants and prostate cancer susceptibility using the iCOGS-genotyping array. *Br J Cancer*. 2016;114(8):945–52. <https://doi.org/10.1038/bjc.2016.50>.
41. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Human Genet*. 2012;91(2):224–37. <https://doi.org/10.1016/j.ajhg.2012.06.007>.
42. Lin Y-C, Hsieh A-R, Hsiao C-L, Wu S-J, Wang H-M, Lian I-B, et al. Identifying rare and common disease associated variants in genomic data using Parkinson's disease as a model. *J Biomed Sci*. 2014;21:88. <https://doi.org/10.1186/s12929-014-0088-9>.
43. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet*. 2013;92(6):841–53. <https://doi.org/10.1016/j.ajhg.2013.04.015>.
44. Michels KB, Greenland S, Rosner BA. Does body mass index adequately capture the relation of body composition and body size to health outcomes? *Am J Epidemiol*. 1998;147(2):167–72.
45. Schweder T, Spjøtvoll E. Plots of *P*-values to evaluate many tests simultaneously. *Biometrika*. 1982;69(3):493–502.
46. Mishra L, Shetty K, Tang Y, Stuart A, Byers SW. The role of TGF- $\beta$  and Wnt signaling in gastrointestinal stem cells and cancer. *Oncogene*. 2005;24(37):5775–89. <https://doi.org/10.1038/sj.onc.1208924>.
47. Jiang X, Castela JE, Vandenberg D, Carracedo A, Redondo CM, Conti DV, et al. Genetic variations in *SMAD7* are associated with colorectal cancer risk in the Colon Cancer family registry. *PLoS One*. 2013;8(4):e60464. <https://doi.org/10.1371/journal.pone.0060464>.
48. Schmit SL, Schumacher FR, Edlund CK, Conti DV, Ihenacho U, Wan P, et al. Genome-wide association study of colorectal cancer in Hispanics. *Carcinogenesis*. 2016;37(6):547–56. <https://doi.org/10.1093/carcin/bgw046>.
49. Baas AF, Medic J, van 't Slot R, de Kovel CG, Zhernakova A, Geelkerken RH, et al. Association of the TGF- $\beta$  receptor genes with abdominal aortic aneurysm. *Eur J Hum Genet*. 2010;18(2):240–4. <https://doi.org/10.1038/ejhg.2009.141>.
50. Slattery ML, Herrick JS, Lundgreen A, Wolff RK. Genetic variation in the TGF- $\beta$  signaling pathway and colon and rectal cancer risk. *Cancer Epidemiol Biomark Prev*. 2011;20(1):57–69. <https://doi.org/10.1158/1055-9965.EPI-10-0843>.
51. Slattery ML, Lundgreen A, Herrick JS, Wolff RK, Caan BJ. Genetic variation in the transforming growth factor- $\beta$  signaling pathway and survival after diagnosis with colon and rectal cancer. *Cancer*. 2011;117(18):4175–83. <https://doi.org/10.1002/cncr.26018>.
52. The National Center for Biotechnology Information (NCBI). NCBI 1000 Genomes Browser, Phase 3. <https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>. Accessed 4 July 2018.
53. Robertson DS, Prevost AT, Bowden J. Accounting for selection and correlation in the analysis of two-stage genome-wide association studies. *Biostatistics*. 2016;17(4):634–49. <https://doi.org/10.1093/biostatistics/kxw012>.
54. Family L, Bensen JT, Troester MA, Wu MC, Anders CK, Olshan AF. Single-nucleotide polymorphisms in DNA bypass polymerase genes and association with breast cancer and breast cancer subtypes among African Americans and whites. *Breast Cancer Res Treat*. 2015;149(1):181–90. <https://doi.org/10.1007/s10549-014-3203-4>.
55. Tang W, Tang J, Zhao Y, Qin Y, Jin G, Xu X, et al. Exome-wide association study identified new risk loci for Hirschsprung's disease. *Mol Neurobiol*. 2017;54(3):1777–85. <https://doi.org/10.1007/s12035-016-9752-2>.
56. Hasegawa T, Kojima K, Kawai Y, Misawa K, Mimori T, Nagasaki M. AP-SKAT: highly-efficient genome-wide rare variant association test. *BMC Genomics*. 2016;17(1):745. <https://doi.org/10.1186/s12864-016-3094-3>.
57. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWAS: illuminating the dark road from association to function. *Am J Hum Genet*. 2013;93(5):779–97. <https://doi.org/10.1016/j.ajhg.2013.10.012>.
58. Spencer AV, Cox A, Lin W-Y, Easton DF, Michailidou K, Walters K. Incorporating functional genomic information in genetic association studies using an empirical Bayes approach. *Genet Epidemiol*. 2016;40(3):176–87. <https://doi.org/10.1002/gepi.21956>.
59. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet*. 2008;17(2):R143–50. <https://doi.org/10.1093/hmg/ddn268>.
60. Koboldt DC, Kanchi KL, Gui B, Larson DE, Fulton RS, Isaacs WB, et al. Rare variation in *TET2* is associated with clinically relevant prostate carcinoma in African Americans. *Cancer Epidemiol Biomark Prev*. 2016;25(11):1456–63. <https://doi.org/10.1158/1055-9965.EPI-16-0373>.
61. Wang H, Taverna D, Stram DO, Fortini BK, Cheng I, Wilkens LR, et al. Genetic variation in the inflammation and innate immunity pathways and colorectal cancer risk. *Cancer Epidemiol Biomark Prev*. 2013;22(11):2094–101. <https://doi.org/10.1158/1055-9965.EPI-13-0694>.
62. Pudova EA, Kudryavtseva AV, Fedorova MS, Zaretsky AR, Shcherbo DS, Lukyanova EN, et al. HK3 overexpression associated with epithelial-mesenchymal transition in colorectal cancer. *BMC Genomics*. 2018;19(Suppl 3):113. <https://doi.org/10.1186/s12864-018-4477-4>.
63. Thingholm LB, Andersen L, Makalic E, Southey MC, Thomassen M, Hansen LL. Strategies for integrated analysis of genetic, epigenetic, and gene expression variation in cancer: addressing the challenges. *Front Genet*. 2016;7:2. <https://doi.org/10.3389/fgene.2016.00002>.
64. Larson NB, McDonnell S, Albright LC, Teerlink C, Stanford J, Ostrander EA, et al. gsSKAT: rapid gene set analysis and multiple testing correction for rare-variant association studies using weighted linear kernels. *Genet Epidemiol*. 2017;41(4):297–308. <https://doi.org/10.1002/gepi.22036>.
65. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*. 2012;336(6082):740–3. <https://doi.org/10.1126/science.1217283>.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

