Received: 9 June 2020 | Revised: 10 October 2020 | Accepted: 9 November 2020

DOI: 10.1002/gepi.22372

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

BRIEF REPORT

Statistical methods with exhaustive search in the identification of gene–gene interactions for colorectal cancer

Somayeh kafaie[1] | **Ling Xu**[1] | **Ting Hu**[1,2]

[1]Department of Computer Science, Memorial University, St. John's, Newfoundland, Canada

[2]School of Computing, Queen's University, Kingston, Ontario, Canada

Correspondence
Ting Hu, School of Computing, Queen's University, Kingston, Ontario, Canada.
Email: ting.hu@queensu.ca

Funding information
Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: Discovery Grant RGPIN-2016-04699

Abstract

Though additive forms of heritability are primarily studied in genetics, nonlinear, non-additive gene–gene interactions, that is, *epistasis*, could explain a portion of the missing heritability in complex human diseases including cancer. In recent years, powerful computational methods have been introduced to understand multivariable genetic factors of these complex human diseases in extremely high-dimensional genome-wide data. In this study, we investigated the performance of three powerful methods, BOolean Operation-based Screening and Testing (BOOST), FastEpistasis, and Tree-based Epistasis Association Mapping (TEAM) to identify interacting genetic risk factors of colorectal cancer (CRC) for genome-wide association studies (GWAS). After quality-control based data preprocessing, we applied these three algorithms to a CRC GWAS data set, and selected the top-ranked 100 single-nucleotide polymorphism (SNP) pairs identified by each method (251 SNPs in total), among which 74 pairs were common between FastEpistasis and BOOST. The identified SNPs by BOOST, FastEpistasis, and TEAM mapped to 58, 57, and 62 genes, respectively. Some genes highlighted by our study, including *MACF1*, *USP49*, *SMAD2*, *SMAD3*, *TGFBR1*, and *RHOA*, have been detected in previous CRC-related research. We also identified some new genes with potential biological relevance to CRC such as *CCDC32*. Furthermore, we constructed the network of these top SNP pairs for three methods, and the patterns identified in the networks show that some SNPs including rs2412531, rs349699, and rs17142011 play a crucial role in the classification of disease status in our study.

KEYWORDS
colorectal cancer, complex diseases, epistasis, gene–gene interaction, genome-wide association studies

1 | INTRODUCTION

For genome-wide association studies (GWAS), the most common objective is to identify genetic regions (loci) and their associated phenotypic traits (including diseases). In recent years, successes have been reported in using GWAS to identify genetic variations that contribute to the risk of prostate cancer (Eeles et al., 2013), Parkinson's disease (Martins et al., 2011), type 2 diabetes (Billings & Florez, 2010), obesity (Loos & Yeo, 2014), Crohn's disease (Franke et al., 2010), and heart disorders (DenHoed et al., 2013).

In fact, a series of statistical and computational methods have been developed so far to provide insights into genetic variants associated with complex diseases. However, many studies apply the uni-variable approach, where a single genetic variant is being scored and ranked based on the significance of its association with the disease phenotype, mostly due to the extremely high dimensionality of GWAS data which could include up to a million variables. These initial GWAS analyses, that have adopted single-nucleotide polymorphism (SNP)-based methods, provide stepping stones for further exploration, but usually have fallen short for discovering new variants explaining disease heritability effectively (Manolio et al., 2009; Szymczak et al., 2009).

Recent research has seen a trend of developing new algorithms for detecting gene–gene interactions rather than focusing only on the single-locus contribution to phenotypic variations. The nonlinear, non-additive gene–gene interaction, also called *epistasis*, is considered to account for some of the missing heritability, which is ignored and unaccounted for in uni-variable approaches (Collins et al., 2013; Cordell, 2002; Dorani et al., 2018; Hu, Chen, Kiralis, Collins, et al., 2013; Hu, Chen, Kiralis, & Moore, 2013; Hu et al., 2011; Kafaie et al., 2019; Moore, 2003; Raghavan & Tosto, 2017; Schubert et al., 2019). The multivariable methods attempt to analyze all possible SNP combinations of various degrees and capture the most significant contributing ones. However, traditional parametric statistical methods such as linear and logistic regression cannot be easily applied to epistasis detection because of the sparseness of GWAS data in extreme high dimensions. Also, the burden of high computational cost adds to the difficulty of developing powerful epistasis detection algorithms (Manduchi et al., 2018).

To detect gene–gene interactions, in this study, we investigated three algorithms, BOolean Operation-based Screening and Testing (BOOST) (Wan et al., 2010), FastEpistasis (Schüpbach et al., 2010), and Tree-based Epistasis Association Mapping (TEAM) (Zhang et al., 2010). We applied these three powerful algorithms to identify interacting genetic risk factors of colorectal cancer (CRC), which is the third most common cancer globally (Jemal et al., 2011), and caused more than 50,000 deaths in the United States alone in 2018 (Siegel et al., 2018). Previous GWAS have identified one SNP (rs6983267) at 8q24.21 ($p = 1.72 \times 10^{-7}$, allelic test) (Tomlinson et al., 2007) and three SNPs in SMAD7 (involved in TGF-β and Wnt signaling) strongly associated with CRC (Broderick et al., 2007).

In our study, the GWAS data set was collected from the population in the Canadian province of Newfoundland and Labrador. All three algorithms were able to rank two-way SNP interactions based on their association with the disease, and identified 251 SNPs, in total, among two-way interaction pairs as having the most significant impact. The functional enrichment analysis on protein products produced by the identified SNPs also indicated significant biological pathways for future references.

2 | METHODOLOGY

2.1 | Overview

Figure 1 shows the flow of data pre-processing and data analyzing procedures. The first step was to filter out inadequate markers and samples within the original data set; hence the original data set containing 265,181 SNPs and 1352 samples was preprocessed into a clean data set of 253,657 SNPs and 1060 samples. Then, we applied the three algorithms, BOOST, FastEpistasis, and TEAM, to the preprocessed data set, and performed functional enrichment analysis on the identified SNP pairs using the tool Database for Annotation, Visualization and Integrated Discovery (DAVID) (Jiao et al., 2012).

2.2 | CRC GWAS data preprocessing

The case–control data set used in our study was sampled from the population in the province of Newfoundland and Labrador, Canada. The Colorectal Transdisciplinary (CORECT) Study coordinated the sampling of the aforementioned data set. It has been ensured that CRC patients and healthy participants share statistically similar geographical, sex, and age compositions. The data set was acquired by genotyping two pegs with a custom Affymetrix genome-wide platform (the Axiom CORECT set). It contained 265,181 SNPs and 1352 samples without duplicates, among which 856 were cases and 496 were controls (811 males and 541 females). To extract the most amount of information from and facilitate subsequent processing of the data set, PLINK (Purcell et al., 2007), a whole genome data analysis toolset, was used.

The next two steps were per-individual and per-marker quality control (QC) procedures. During the per-individual

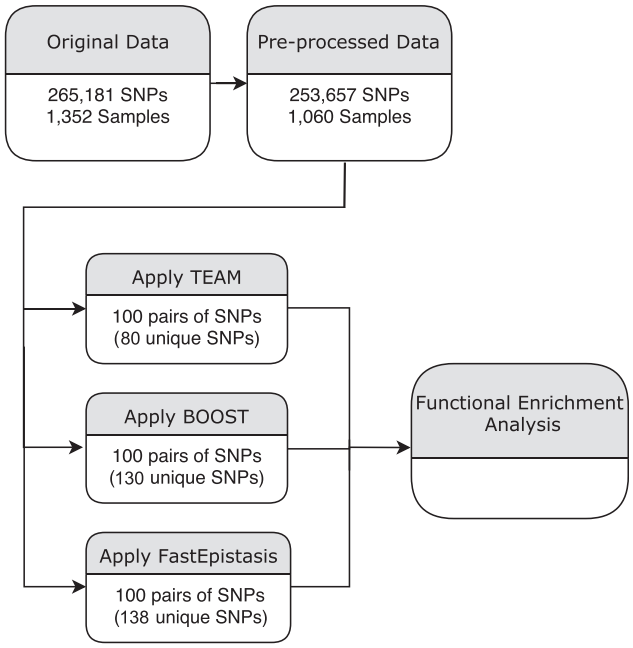

FIGURE 1 Flow-chart for data pre-processing and data analyzing procedures. First, the original data set containing 265,181 SNPs and 1352 samples was cleaned up into a preprocessed data set of 253,657 SNPs and 1060 samples. Next, three exhaustive algorithms, namely BOOST, FastEpistasis, and TEAM were applied to the clean data set. Last, we performed functional enrichment analysis on the three identified SNP sets by the mentioned algorithms. BOOST, BOolean Operation-based Screening and Testing; SNP, single-nucleotide polymorphism; TEAM, Tree-based Epistasis Association Mapping

QC phase, the following groups of individuals were identified: 13 samples with discordant sex information, 24 samples with outlying heterozygosity rate ($\pm 3 \times SD$), 3 samples with elevated missing genotype data rate (≥ 0.03), 230 duplicated or related samples with IBD[1] > 0.185, and 53 samples of divergent ancestry with a second principal component score less than 0.072 through principal component analysis. During the per-individual QC phase, 292 individuals failed to pass the above filters.

During the per-marker QC phase, the following categories of individuals were filtered out: 6,057 sex chromosome SNPs, 174 SNPs with a missing data rate over 3%, 3782 SNPs with different genotype call rates between cases and controls, 168 SNPs with Hardy–Weinberg equilibrium values greater than 0.00001, and 1343 SNPs with minor allele frequencies (MAFs) less than 5%. In total, 11,524 SNPs were removed from the data set. Then missing genotypes were imputed in proportion to the ratio of dominant alleles against recessive alleles. The data set after per-individual and per-marker quality control procedures had 253,657 SNPs and 1060 individuals, among which 603 were cases and 457 were controls (630 males and 430 females).

2.3 | BOOST, FastEpistasis, and TEAM

The goal of our study is to identify gene–gene interactions from GWAS of the CRC data using the following three exhaustive search methods and investigate their performances. A simple but powerful method, named "BOOST" (Wan et al., 2010), allows examination of all pairwise interactions in genomic case-control studies in a remarkably fast manner. BOOST can be used for the purpose of discovering unknown gene–gene interactions that underlie complex diseases, such as CRC. The implementation of the method includes two phases: screening and testing. During the screening stage, all pairwise interactions are evaluated by the Kirkwood superposition approximation (KSA) and the significant interactions will be guaranteed to enter the testing phase. During the testing stage, the classical likelihood ratio test is employed to measure the interaction effects of selected SNP pairs.

FastEpistasis is implemented in PLINK (option: fast-epistasis) and is used as a fast method to test for interactions (Schüpbach et al., 2010). It is based on a 2×2 contingency table of allele counts and tests an SNP pair for epistasis by comparing their linkage disequilibrium (LD) in cases and controls.

TEAM is another efficient algorithm which significantly speeds up epistasis detection for GWAS (Zhang et al., 2010). It is an exhaustive method by utilizing the minimum spanning tree structure to incrementally updates the contingency tables for epistatic tests without scanning all individuals. The algorithm supports any statistical tests that makes use of contingency tables; therefore has broader applicability and is more efficient than most existing algorithms for GWAS.

We chose these three approaches because (1) BOOST, FastEpistasis, and TEAM all are well-known tools for detecting pairwise interactions that use the exhaustive search strategy and their packages and manuals are readily available online; (2) they have been used and recommended in studies and reviews as powerful and computationally efficient tools for scanning epistatic interactions on GWAS (Cowman & Koyutürk, 2017; Guo et al., 2014; Murk & DeWan, 2016; Wang et al., 2011). We generally set the parameters of these methods as default, and 100 was chosen as the number of permutations for TEAM.

[1]IBD (identity by descent) or the degree of recent shared ancestry for a pair of individuals can be estimated using genome-wide IBS (identical by state) data. The expectation is that $IBD = 1$ for duplicates or monozygotic twins, $IBD = 0.5$ for first-degree relatives, $IBD = 0.25$ for second-degree relatives, and $IBD = 0.125$ for third-degree relatives.

2.4 | Functional enrichment analysis

Utilizing online resources including ENSEMBL[2] and the National Center for Biotechnology Information (NCBI)[3] databases, the identified top ranking SNPs were annotated with functional information. NCBI and ENSEMBL provide biological information on the allele, chromosome, and gene information for each SNP.

Then, a functional enrichment analysis was conducted on the top-ranked SNPs discovered in the three gene–gene interaction detecting algorithms via the Database for Annotation, Visualization, and Integrated Discovery (DAVID) bioinformatics tool (Jiao et al., 2012).

3 | RESULTS

In this section, we first show the application results of the three mentioned methods on the CRC GWAS data set and then the functional annotation chart produced by DAVID. We also apply machine learning techniques to verify the significance of top SNPs selected by the methods on predicting CRC.

3.1 | Top-ranked 100 SNP pairs

Among the top 100 ranked SNP pairs detected by FastEpistasis and BOOST, we found 74 common pairs of SNPs. TEAM-detected SNP pairs, on the other hand, were in a separate set, which had no pairs in common with these of FastEpistasis and BOOST. There were 130, 138, and 80 unique SNPs within the top-ranked SNP pairs detected by BOOST, FastEpistasis, and TEAM, respectively (see Figure 1). The similar results from BOOST and FastEpistasis indicate common factors considered in the ranking statistics used in both algorithms, whereas TEAM utilizes a different approach when computing its own ranking statistics.

Figures 2–4 use network diagrams of SNPs in the top 100 ranked SNP pairs identified by BOOST, FastEpistasis, and TEAM to illustrate the interconnections and pairwise-relationships among the depicted SNPs. 92.3% of the SNPs identified by BOOST encode protein products, 88.4% by FastEpistasis, and 75% by TEAM.

The two networks of SNPs identified by BOOST and FastEpistasis showed a similar fragmented structure with 33 and 41 connected components, respectively, and shared 74 pairs of SNPs. Whereas the SNP interaction network identified by TEAM contained only one giant component connected through three hub SNPs.

Boost and FastEpistasis networks also shared an SNP with the most connections to other SNPs, namely rs2412531, which was found to interact with the same set of nine SNPs in both networks. In addition, SNP rs349699 was found to interact with the same set of five SNPs in both networks. Other common high-degree SNPs shared by these two networks include rs6808, rs122800, rs1487324, and rs349699. Echoing the distinct structure showed in the TEAM network, its three hub SNPs, rs17142011, rs6355, and rs2072193 were absent in the other two networks.

3.2 | Enriched gene functional terms

Each of the three sets of SNPs and its coding protein products, identified by BOOST, FastEpistasis, and TEAM respectively, were fed into DAVID for a functional enrichment analysis. Seven functional categories, that is, Disease, GOTERM, Pathway, UP_SEQ_FEATURE, SP_PIR_KEYWORDS, SMART, and INTERPRO were chosen by default settings in DAVID. We set the gene count threshold as two due to the limited amount of functional categories having more than two markers and the p value cutoff as 0.1. The details of enrichment analysis for top SNPs chosen by BOOST, FastEpistasis, and TEAM have been provided in excel files as a supplement.

Tables 1–3 show the most significantly enriched functional categories. For the set of genes identified by BOOST, the most significantly enriched term was "Whey acidic protein-type 4-disulphide core" within the INTERPRO category with 4 out of 58 genes in total and a significance level of 9.46×10^{-6}. For the set of genes identified by FastEpistasis, the most significantly enriched term was "regulation of ERK1 and ERK2 cascade" within the GOTERM_BP_DIRECT category with 3 out of 57 genes in total and a significance level of 1.85×10^{-3}. For the set of genes identified by TEAM, the most significantly enriched term was "temperature homeostasis" within the GOTERM_BP_DIRECT category with 3 out of 62 genes in total and a significance level of 1.06×10^{-3}.

3.3 | Phenotype variation explained by top SNPs

We developed machine learning models to verify the significance of selected SNPs regarding the disease by measuring the accuracy of disease prediction. In fact, from each method, we selected SNPs involved in the strongest pairwise interactions and trained five models with stochastic gradient

[2] http://www.ensembl.org
[3] https://www.ncbi.nlm.nih.gov/

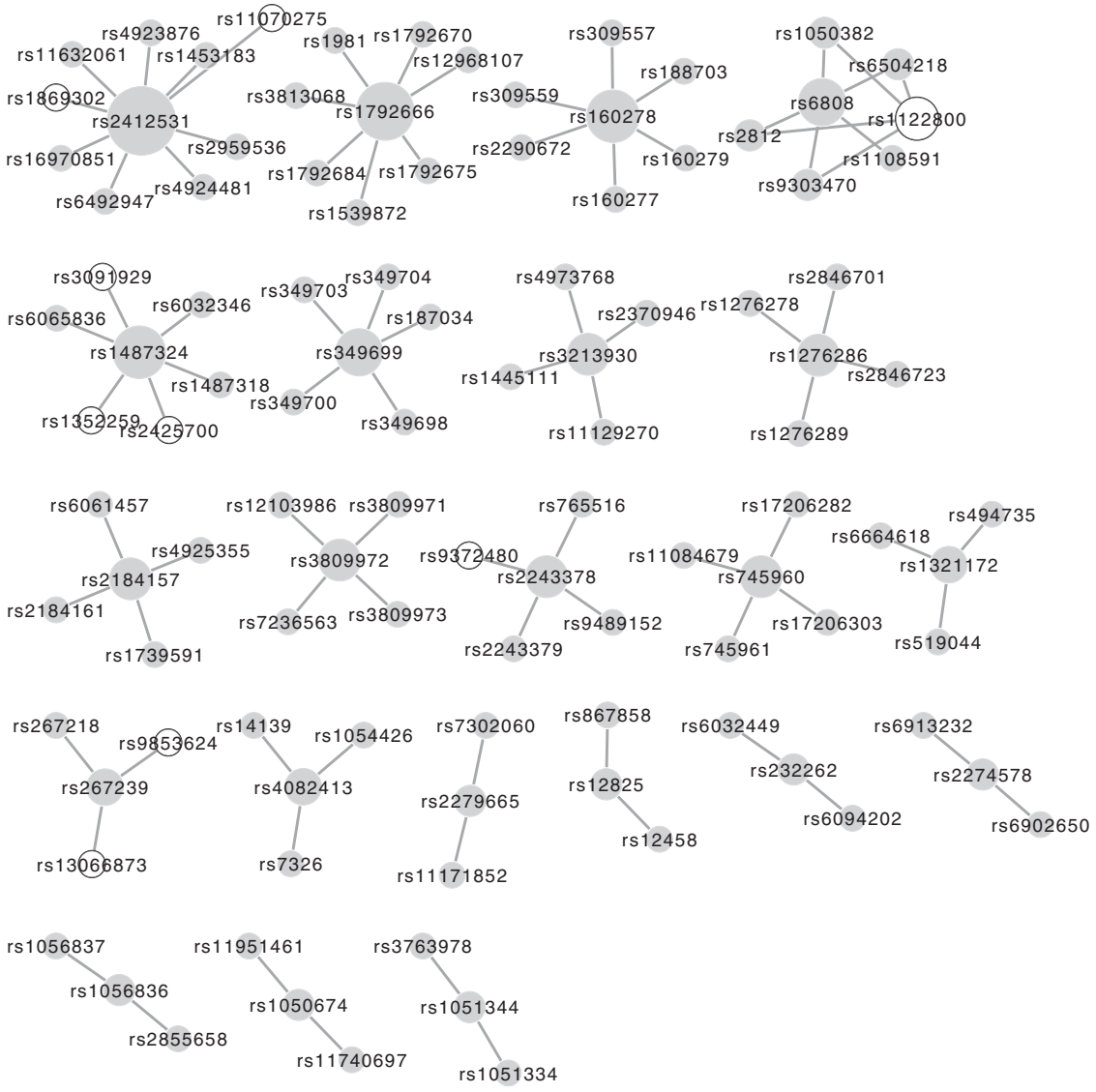

FIGURE 2 A network of SNPs available in the top 100 ranking SNP pairs identified by BOOST. Using Cytoscape Shannon et al. (2003), a software for drawing networks, a network was drawn showing the pairwise relationship between SNPs identified by BOOST. The darker circle represents SNPs encoding protein products whereas the lighter circle represents noncoding SNPs. There were 130 unique SNPs in total, among which 10 were noncoding and 120 were coding SNPs. Isolated SNP pairs were not included in this figure. BOOST, BOolean Operation-based Screening and Testing; SNP, single-nucleotide polymorphism

descent (SGD), K nearest neighbors (KNN), support vector machine (SVM), random forest, and multilayer perceptron (MLP). Since there were 80 unique SNPs within the top 100 ranking pairs of SNPs detected by TEAM, we selected the top 80 SNPs identified by each method as well as allelic odds ratio (Clarke et al., 2011) as the explanatory features. The models' hyper-parameters were optimized by applying a k-fold cross-validation with random search. We used 90% of the data set for training, and the other 10% was used as the test set to measure the accuracy. The average accuracy calculated for different approaches has been shown in Figures 5.

As shown in the figure, the trend among different approaches and their comparative performance was almost the same across all machine learning models. The SNPs identified by TEAM predicted the disease with higher accuracy than the SNPs selected by other approaches in almost all models. Furthermore, while TEAM and FastEpistasis outperformed allelic odds ratio in all cases, the performance of BOOST and allelic odds ratio was comparable in many cases. Also, these results show that the SNPs identified by BOOST cannot explain the phenotype variation as efficiently as TEAM and FastEpistasis. As explained before, One reason could be that in our model we

FIGURE 3 A network of SNPs available in the top 100 ranking SNP pairs identified by FastEpistasis. Using Cytoscape, a software for drawing networks, a network was drawn showing the pairwise relationship between SNPs identified by FastEpistasis. The darker circle represents SNPs encoding protein products whereas the lighter circle represents noncoding SNPs. There were 138 unique SNPs in total, among which 16 were noncoding and 122 were coding SNPs. Isolated SNP pairs were not included in this figure. SNP, SNP, single-nucleotide polymorphism

used individual SNPs as the features while BOOST does not consider the main effect of SNPs and is looking only for SNPS with strong interactions (Guo et al., 2014; Wang et al., 2011).

4 | DISCUSSION

Following the speculation of the intricacies of complex human diseases, the classification factors of disease phenotypes are more likely to be gene–gene interactions instead of genes contributing individually to the disease (Cordell, 2009). Epistasis detection therefore can potentially enrich our existing knowledge of the disease etiology and help understand the genetic architecture of the diseases. Powerful statistical and computational algorithms have shown success for mining extremely high-dimensional data in academic disciplines of finance and engineering, and have started to see potentials in applications to analyzing biomedical data in recent years.

In this study, we explored the applications of three exhaustive pairwise epistasis detection algorithms, BOOST, FastEpistasis, and TEAM in identifying SNP–SNP interactions associated with CRC. After data preprocessing, the three algorithms were applied to exhaustively evaluate all pairwise interactions of 253,657 SNPs in a CRC GWAS data set. Using a computer workstation running Ubuntu 20.04 and having 3.5 GHZ Intel Xeon Quad-Core Processor and 128 GB RAM, BOOST completed the task in about 5 h. On the other hand, TEAM had the longest run-time which was

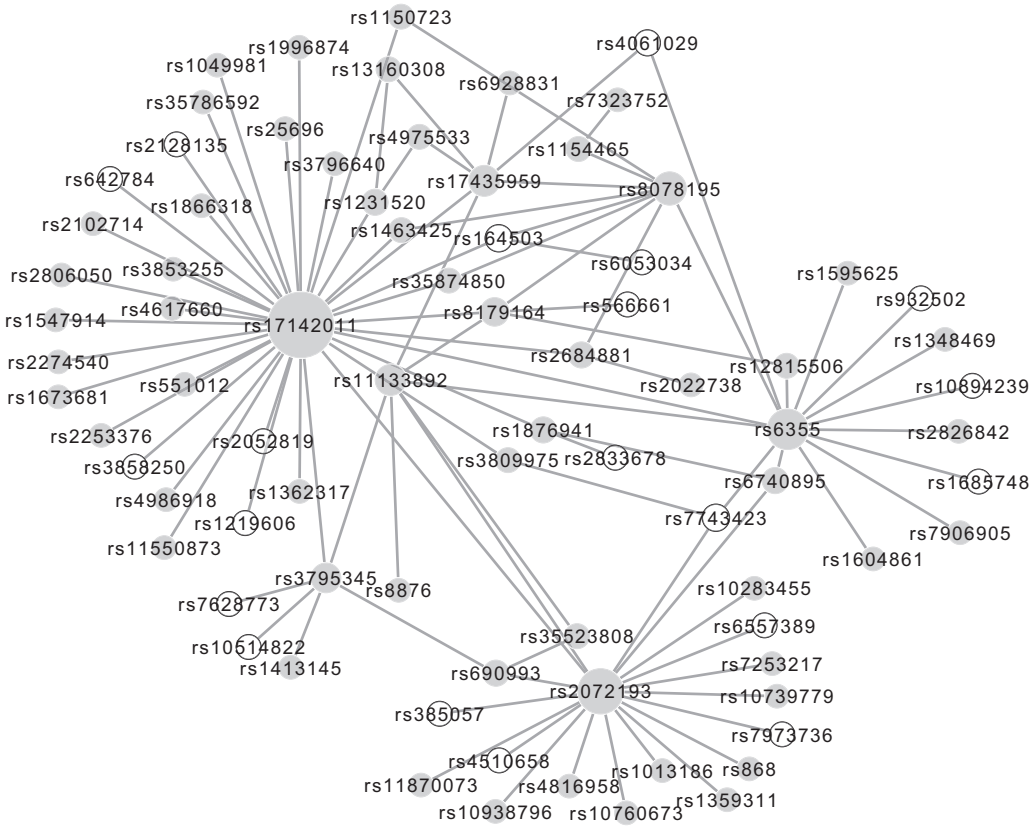

FIGURE 4 A network of SNPs available in the top 100 ranking SNP pairs identified by TEAM. Using Cytoscape, a software for drawing networks, a network was drawn showing the pairwise relationship between SNPs identified by TEAM. The darker circle represents SNPs encoding protein products whereas the lighter circle represents noncoding SNPs. There were 80 unique SNPs in total, among which 20 were noncoding and 60 were coding SNPs. SNP, SNP, single-nucleotide polymorphism; TEAM, TEAM, Tree-based Epistasis Association Mapping

about 11.5 days (271 h), and the experiment was completed by FastEpistasis in around 4 days (98.5 h).

We then closely investigated the set of 100 top SNP pairs ranked by the three algorithms based on the importance of their contribution to the disease phenotypic association. BOOST and FastEpistasis shared 74 pairs of SNPs, while the set of SNP pairs chosen by TEAM was completely different. Their difference in performance can be due to the fact that they define the interaction effect differently. Although TEAM performs best on data with main effect, BOOST identifies statistical interactions without considering the main effect (Guo et al., 2014; Wang et al., 2011). In fact, by applying χ^2 test, TEAM finds pairwise interactions without any assumption about the data. On the other hand, BOOST uses a log likelihood ratio test and has a better performance when interaction effect contributes significantly to the model (Wang et al., 2011).

The three coding SNP sets, discovered by BOOST, FastEpistasis, and TEAM, mapped to 58, 57, and 62 genes, respectively. A number of these CRC-associated genes have already been discovered in the literature, including *MACF1*, *USP49*, *SMAD2*, *SMAD3*, *TGFBR1*, and *RHOA*. Microtubule actin crosslinking factor 1 (*MACF1*) was identified as one of the oncofetal biomarkers by protein profiling of serum samples from CRC patients, healthy control adults, and fetus (Ma et al., 2012). *USP49* was found that can increase cell sensitivity to etoposide (Eto)-induced DNA damage and was suggested as a tumor suppressor during the development of CRC (Tu et al. (2018)). Some analyses indicate that failure to express *SMAD2* is associated with advanced-stage disease, poor prognosis, and shorter survival (W. Xie et al., 2003). (Fleming et al., 2013) suggested *SMAD2* and *SMAD3* mutations as true contributors to the mutation burden in CRCs, and the result of (Zhu et al., 2017) showed that *SMAD3* mutant mice develop colon cancer with overexpression of COX-2. Many studies indicated Ras homologue family member A (*RHOA*) as a tumor suppressor in CRC and showed that low *RHOA* expression is associated with poor prognosis and significantly shorter survival (Arango et al., 2005; Dopeso et al., 2018; Jeong et al., 2016; Rodrigues et al., 2014). Furthermore, Transforming growth factor beta receptor type 1 (*TGFBR1*) has been found to contribute to the

TABLE 1 Enrichment gene oncology (GO) terms on 58 BOOST-identified genes

Category	Term	Protein count	*p* Value	Bonferroni	Benjamini	FDR
INTERPRO	Whey acidic protein-type 4-disulphide core	4	9.46×10^{-6}	1.32×10^{-3}	1.32×10^{-3}	1.11×10^{-2}
UP_KEYWORDS	Serine protease inhibitor	5	5.65×10^{-5}	7.26×10^{-3}	7.26×10^{-3}	6.57×10^{-2}
GOTERM_MF_DIRECT	Serine-type endopeptidase inhibitor activity	5	1.27×10^{-4}	1.37×10^{-2}	1.37×10^{-2}	1.43×10^{-1}
UP_KEYWORDS	Protease inhibitor	5	2.33×10^{-4}	2.96×10^{-2}	1.49×10^{-2}	2.71×10^{-1}
GOTERM_BP_DIRECT	Negative regulation of endopeptidase activity	5	2.79×10^{-4}	7.38×10^{-2}	7.38×10^{-2}	3.67×10^{-1}
GOTERM_BP_DIRECT	Dermatan sulfate biosynthetic process	3	4.36×10^{-4}	1.13×10^{-1}	5.81×10^{-2}	5.73×10^{-1}
KEGG_PATHWAY	Hippo signaling pathway	4	4.71×10^{-3}	1.87×10^{-1}	1.87×10^{-1}	4.32
INTERPRO	Uncharacterized protein family; WAP four-disulphide core	2	4.95×10^{-3}	5×10^{-1}	2.93×10^{-1}	5.69
UP_SEQ_FEATURE	domain:WAP 3	2	4.98×10^{-3}	7.04×10^{-1}	7.04×10^{-1}	6.25
UP_SEQ_FEATURE	domain:WAP 2	2	1.24×10^{-2}	9.52×10^{-1}	7.82×10^{-1}	14.91
UP_SEQ_FEATURE	domain:WAP 1	2	1.24×10^{-3}	9.52×10^{-1}	7.82×10^{-1}	14.91
GOTERM_BP_DIRECT	Embryonic foregut morphogenesis	2	2.33×10^{-2}	9.98×10^{-1}	8.85×10^{-1}	26.76
UP_SEQ_FEATURE	domain:BPTI/Kunitz inhibitor	2	2.71×10^{-2}	9.99×10^{-1}	8.93×10^{-1}	29.9
UP_SEQ_FEATURE	domain:WAP	2	2.71×10^{-2}	9.99×10^{-1}	8.93×10^{-1}	29.9
SMART	SM00217:WAP	2	2.89×10^{-2}	6.51×10^{-1}	6.51×10^{-1}	22.99
GOTERM_BP_DIRECT	Regulation of cardiac muscle cell contraction	2	3.1×10^{-2}	9.99×10^{-1}	8.85×10^{-1}	33.98
GOTERM_MF_DIRECT	co-SMAD binding	2	3.15×10^{-2}	9.69×10^{-1}	8.23×10^{-1}	30.29
UP_KEYWORDS	Ehlers-Danlos syndrome	2	3.35×10^{-2}	9.88×10^{-1}	7.69×10^{-1}	32.71
UP_KEYWORDS	Polymorphism	37	3.43×10^{-2}	9.89×10^{-1}	6.76×10^{-1}	33.41
SMART	KU	2	3.69×10^{-2}	7.42×10^{-1}	4.92×10^{-1}	28.53

Abbreviations: BOOST, BOolean Operation-based Screening and Testing; FDR, false discovery rate.

TABLE 2 Enrichment gene oncology (GO) terms on 57 FastEpistasis-identified genes

Category	Term	Protein count	*p* value	Bonferroni	Benjamini	FDR
GOTERM_BP_DIRECT	Regulation of ERK1 and ERK2 cascade	3	1.85×10^{-3}	4.18×10^{-1}	4.18×10^{-1}	2.43
INTERPRO	Uncharacterized protein family, WAP four-disulphide core	2	4.95×10^{-3}	5.39×10^{-1}	5.39×10^{-1}	5.79
KEGG_PATHWAY	Chemical carcinogenesis	3	1.27×10^{-2}	4.38×10^{-1}	4.38×10^{-1}	11.34
UP_KEYWORDS	Polymorphism	36	2.79×10^{-2}	9.69×10^{-1}	9.69×10^{-1}	27.85
GOTERM_BP_DIRECT	Protein phosphorylation	5	2.87×10^{-2}	9.99×10^{-1}	9.86×10^{-1}	32.19
GOTERM_BP_DIRECT	Cell adhesion	5	2.94×10^{-2}	9.99×10^{-1}	9.45×10^{-1}	32.75
GOTERM_BP_DIRECT	Dermatan sulfate biosynthetic process	2	3.03×10^{-2}	9.99×10^{-1}	8.94×10^{-1}	33.61
GOTERM_MF_DIRECT	Enhancer binding	2	3.35×10^{-2}	9.86×10^{-1}	9.86×10^{-1}	32.55
KEGG_PATHWAY	Cell adhesion molecules (CAMs)	3	3.73×10^{-2}	8.19×10^{-1}	5.74×10^{-1}	30.03

Abbreviations: FDR, false discovery rate.

TABLE 3 Enrichment gene oncology (GO) terms on 62 TEAM-identified genes

Category	Term	Protein Count	*p* value	Bonferroni	Benjamini	FDR
GOTERM_BP_DIRECT	Temperature homeostasis	3	1.06×10^{-3}	4.51×10^{-1}	4.51×10^{-1}	1.55
GOTERM_BP_DIRECT	Transforming growth factor beta receptor signaling pathway	4	1.68×10^{-3}	6.11×10^{-1}	3.76×10^{-1}	2.43
KEGG_PATHWAY	TGF-beta signaling pathway	4	1.72×10^{-3}	1.24×10^{-1}	1.24×10^{-1}	1.8
UP_KEYWORDS	Oxidoreductase	7	2×10^{-3}	2.58×10^{-1}	2.58×10^{-1}	2.36
GOTERM_BP_DIRECT	Positive regulation of cytokinesis	3	4.03×10^{-3}	8.97×10^{-1}	5.31×10^{-1}	5.74
GOTERM_BP_DIRECT	Ureteric bud development	3	4.25×10^{-3}	9.09×10^{-1}	4.51×10^{-1}	6.04
GOTERM_MF_DIRECT	Positive regulation of stress fiber assembly	3	5.17×10^{-3}	9.46×10^{-1}	4.42×10^{-1}	7.3
UP_KEYWORDS	Cell projection	7	5.72×10^{-3}	5.75×10^{-1}	3.48×10^{-1}	6.62
UP_KEYWORDS	Polymorphism	37	7.45×10^{-3}	6.72×10^{-1}	3.1×10^{-1}	8.53
GOTERM_BP_DIRECT	Regulation of potassium ion transport	2	1.02×10^{-2}	9.97×10^{-1}	6.18×10^{-1}	13.94
UP_KEYWORDS	Phosphoprotein	28	1.19×10^{-2}	8.31×10^{-1}	3.59×10^{-1}	13.26
GOTERM_MF_DIRECT	Transforming growth factor beta receptor, pathway-specific cytoplasmic mediator activity	2	1.3×10^{-2}	8.71×10^{-1}	8.71×10^{-1}	14.54
KEGG_PATHWAY	Colorectal cancer	3	1.37×10^{-2}	6.54×10^{-1}	4.12×10^{-1}	13.53
UP_SEQ_FEATURE	Nucleotide phosphate-binding region	3	1.51×10^{-2}	9.84×10^{-1}	9.84×10^{-1}	18.16
UP_KEYWORDS	Alternative splicing	33	1.54×10^{-2}	9.01×10^{-1}	3.71×10^{-1}	16.93
GOTERM_CC_DIRECT	Endoplasmic reticulum	7	1.55×10^{-2}	8.49×10^{-1}	8.49×10^{-1}	16.46
KEGG_PATHWAY	Adherens junction	3	1.77×10^{-2}	7.47×10^{-1}	3.68×10^{-1}	17.17
GOTERM_CC_DIRECT	Endosome	4	1.8×10^{-2}	8.89×10^{-1}	6.67×10^{-1}	18.85
UP_SEQ_FEATURE	domain:MH1	2	1.86×10^{-2}	9.94×10^{-1}	9.21×10^{-1}	21.88
UP_SEQ_FEATURE	domain:MH2	2	1.86×10^{-2}	9.94×10^{-1}	9.21×10^{-1}	21.88
SMART	SM00524:DWB	2	1.89×10^{-2}	5.43×10^{-1}	5.43×10^{-1}	16.13
GOTERM_CC_DIRECT	SMAD protein complex	2	1.96×10^{-2}	9.09×10^{-1}	5.49×10^{-1}	20.34
INTERPRO	SMAD domain, Dwarfin-type	2	1.97×10^{-2}	9.66×10^{-1}	9.66×10^{-1}	21.51
INTERPRO	MAD homology, MH1	2	1.97×10^{-2}	9.66×10^{-1}	9.66×10^{-1}	21.51
INTERPRO	Dwarfin	2	1.97×10^{-2}	9.66×10^{-1}	9.66×10^{-1}	21.51
UP_KEYWORDS	Aortic aneurysm	2	2.48×10^{-2}	9.76×10^{-1}	4.65×10^{-1}	25.92
UP_SEQ_FEATURE	Sequence variant	37	2.52×10^{-2}	9.99×10^{-1}	9.01×10^{-1}	28.56
GOTERM_BP_DIRECT	Negative regulation of cytosolic calcium ion concentration	2	2.53×10^{-2}	9.99×10^{-1}	8.73×10^{-1}	31.3
GOTERM_BP_DIRECT	Axonogenesis	3	2.61×10^{-2}	9.99×10^{-1}	8.44×10^{-1}	32.07
SMART	DWA	2	2.83×10^{-2}	6.91×10^{-1}	4.45×10^{-1}	23.2
GOTERM_BP_DIRECT	Positive regulation of gene expression	4	2.93×10^{-2}	9.99×10^{-1}	8.44×10^{-1}	35.3
INTERPRO	MAD homology 1, Dwarfin-type	2	2.93×10^{-2}	9.94×10^{-1}	9.2×10^{-1}	30.47
GOTERM_BP_DIRECT	Positive regulation of catenin import into nucleus	2	3.03×10^{-2}	9.99×10^{-1}	8.23×10^{-1}	36.27
GOTERM_BP_DIRECT	Ethanol oxidation	2	3.03×10^{-2}	9.99×10^{-1}	8.23×10^{-1}	36.28
GOTERM_BP_DIRECT	Roundabout signaling pathway	2	3.03×10^{-2}	9.99×10^{-1}	8.23×10^{-1}	36.28

TABLE 3 (Continued)

Category	Term	Protein Count	*p* value	Bonferroni	Benjamini	FDR
GOTERM_CC_DIRECT	Node of Ranvier	2	3.64×10^{-2}	9.8×10^{-1}	6.23×10^{-1}	30.97
INTERPRO	SMAD domain-like	2	3.89×10^{-2}	9.99×10^{-1}	8.18×10^{-1}	38.62

Abbreviation: FDR, false discovery rate; TEAM, Tree-based Epistasis Association Mapping.

CRC development, and increased tumor cell proliferation (Zeng et al., 2009; R. Zhou et al., 2018).

We also identified some novel SNPs associated with CRC in our results. For instance, SNP rs2412531 which was the largest hub in both BOOST and FastEpistatsis had significant interactions with three SNPs, rs4924481, rs11632061, and rs4923876. In fact, our analyses based on both BOOST and FastEpistasis marked these three interactions (i.e., [rs2412531, rs4924481], [rs2412531, rs11632061], and [rs2412531, rs4923876]) as the three strongest interactions available. Interestingly, all these four SNPs are mapped to the same gene, Coiled-Coil Domain-Containing Protein 32 (*CCDC32*). Although to the best of our knowledge, no research has discussed the association of *CCDC32* with CRC, our study suggested that *CCDC32* could be a potential risk factor worth further experimental investigation.

The pathway term "Colorectal cancer" was significantly enriched ($p = 1.37 \times 10^{-2}$) in top genes found by TEAM. Three genes were included in this category, *TGFBR1*, *RHOA*, and *SMAD3*, which were discussed in a previous paragraph. In addition, terms "Serine protease inhibitor" and "Serine-type endopeptidase inhibitor activity" were significantly enriched ($p = 5.65 \times 10^{-5}$ and $p = 1.27 \times 10^{-4}$, respectively) in top genes ranked by BOOST. It was suggested in the literature that the levels of serine protease in colon tissue interstitial fluid and serum can serve as an indicator of CRC progression (Y. Xie et al., 2016). Another interesting finding was the enriched term "Protease inhibitor" ($p = 2.33 \times 10^{-4}$). It has been reported that proteases are implicated in tumor growth and progression, and protease inhibitors could be considered as a potent strategy in cancer therapy (Eatemadi et al., 2017). The enriched term "Regulation of ERK1 and ERK2 cascade" ($p = 1.85 \times 10^{-3}$) was also evidenced by recent discovery of the correlation of ERK/MAPK signaling pathway with proliferation and apoptosis of colon cancer cells (G. Zhou et al., 2019).

In summary, we applied three powerful computational algorithms to identify the synergistic effect of SNP pairs on phenotypic association of CRC. Using BOOST, FastEpistasis, and TEAM, all pairwise SNP interactions were evaluated on a CRC GWAS data set and three sets of 100 top-ranked SNP pairs were further investigated. We constructed three networks using these three sets of SNP pairs and discussed their properties. We also performed functional enrichment analysis on the top identified SNP pairs using DAVID. Although the effect of

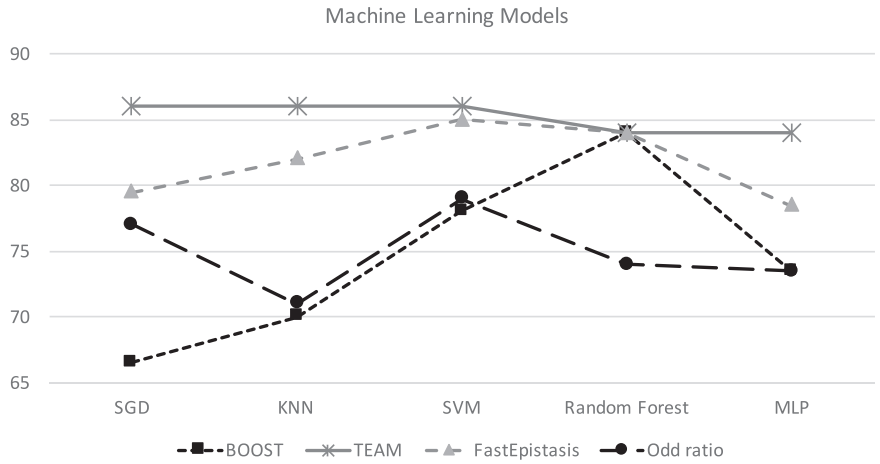

FIGURE 5 The average accuracy of disease prediction using 80 top SNPs as features selected by BOOST, TEAM, FastEpistasis, and allelic odds ratio. We trained five machine learning models, namely logistic regression with SGD, KNN, SVM, random forest, and MLP. BOOST, BOolean Operation-based Screening and Testing; KNN, K nearest neighbors; SGD, stochastic gradient descent; SVM, support vector machine; TEAM, Tree-based Epistasis Association Mapping

some of highlighted SNPs and genes on CRC has already been proved in previous studies, it seems worthy to validate the influence of the rest by further biological research. Identification of these genetic risk factors of CRC can be helpful in identifying people with higher CRC risk, providing more efficient treatments and developing more effective drugs.

ACKNOWLEDGMENTS

This study was supported by the National Sciences and Engineering Research Council (NSERC) of Canada, Discovery Grant RGPIN-2016-04699 to Ting Hu.

DATA AVAILABILITY STATEMENT

The following information was supplied regarding data availability: Colorectal Transdisciplinary (CORECT) Study: https://research.fhcrc.org/peters/en/corect-study.html and https://github.com/MIB-Lab/GeneInteractCRC.

ORCID

Ting Hu http://orcid.org/0000-0001-6382-0602

REFERENCES

Arango, D., Laiho, P., Kokko, A., Alhopuro, P., Sammalkorpi, H., Salovaara, R., Nicorici, D., Hautaniemi, S., Alazzouzi, H., Mecklin, J. K., Järvinen, H., Hemminki, A., Astola, J., Schwartz, S., Jr, & Aaltonen, L. A. (2005). Gene-expression profiling predicts recurrence in dukes' c colorectal cancer. *Gastroenterology*, *129*(3), 874–884.

Billings, L. K., & Florez, J. C. (2010). The genetics of type 2 diabetes: What have we learned from GWAS? *Annals of the New York Academy of Sciences*, *1212*, 59–77.

Broderick, P., Carvajal-Carmona, L., Pittman, A. M., Webb, E., Howarth, K., Rowan, A., Lubbe, S., Spain, S., Sullivan, K., Fielding, S., Jaeger, E. Vijayakrishnan, J., Kemp, Z., Gorman, M., Chandler, I., Papaemmanuil, E., Penegar, S., Wood, W., Sellick, G., ... Houlston, R. S. (2007). A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nature Genetics*, *39*, 1315.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols*, *6*, 121–133.

Collins, R. L., Hu, T., Wejse, C., Sirugo, G., Williams, S. M., & Moore, J. H. (2013). Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis. *BioData Mining*, *6*(1), 4.

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, *11*(20), 2463–2468.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, *10*, 392.

Cowman, T., & Koyutürk, M. (2017). Prioritizing tests of epistasis through hierarchical representation of genomic redundancies. *Nucleic Acids Research*, *45*(14):e131.

DenHoed, M., Eijgelsheim, M., Esko, T., Brundel, B. J., Peal, D. S., Evans, D. M., Nolte, I. J., Segré, A. V., Holm, H. Handsaker, R. E., Westra, H.-J., Johnson, T., Isaacs, A., Yang, J., Lundby, A., Zhao, J. H., Kim, Y. J., Go, M. J., Almgren, P., ... Loss, R. J. F. (2013). Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nature Genetics*, *45*, 621.

Dopeso, H., Rodrigues, P., Bilic, J., Bazzocco, S., Cartón-García, F., Macaya, I., de Marcondes, P. G., Anguita, E., Masanas, M., Jiménez-Flores, L. M., Martínez-Barriocanal, Á., Nieto, R., Segura, M. F., Schwartz Jr, S., Mariadason, J. M., & Arango, D. (2018). Mechanisms of inactivation of the tumour suppressor gene rhoa in colorectal cancer. *British Journal of Cancer*, *118*(1), 106–116.

Dorani, F., Hu, T., Woods, M. O., & Zhai, G. (2018). Ensemble learning for detecting gene-gene interactions in colorectal cancer. *PeerJ*, *6*, e5854.

Eatemadi, A., Aiyelabegan, H. T., Negahdari, B., Mazlomi, M. A., Daraee, H., Daraee, N., Eatemadi, R., & Sadroddiny, E. (2017). Role of protease and protease inhibitors in cancer pathogenesis and treatment. *Biomedicine & Pharmacotherapy*, *86*, 221–231.

Eeles, R. A., AlOlama, A. A., Benlloch, S., Saunders, E. J., Leongamornlert, D. A., Tymrakiewicz, M., Ghoussaini, M., Luccarini, C., Dennis, J., Jugurnauth-Little, S., Dadaev, T., Neal, D. E., Hamdy, F. C., Donovan, J. L., Muir, K., Giles, G. G., Severi, G., Wiklund, F., Gronberg, H., ... Easton, D. F. (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nature Genetics*, *45*, 385.

Fleming, N. I., Jorissen, R. N., Mouradov, D., Christie, M., Sakthianandeswaren, A., Palmieri, M., Day, F., Li, S., Tsui, C., Lipton, L., Desai, J., Jones, I. T., McLaughlin, S., Ward, R. L., Hawkins, N. J., Ruszkiewicz, A. R., Moore, J., Zhu, H.-J., Mariadason, J. M., ... Sieber, O. M. (2013). Smad2, smad3 and smad4 mutations in colorectal cancer. *Cancer Research*, *73*(2), 725-735.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., ... Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, *42*, 1118.

Guo, X., Meng, Y., Yu, N., & Pan, Y. (2014). Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering. *BMC Bioinformatics*, *15*(1), 102.

Hu, T., Chen, Y., Kiralis, J. W., Collins, R. L., Wejse, C., Sirugo, G., Williams, S. M., & Moore, J. H. (2013). An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association*, *20*, 630–636.

Hu, T., Chen, Y., Kiralis, J. W., & Moore, J. H. (2013). ViSEN: Methodology and software for visualization of statistical epistasis networks. *Genetic Epidemiology*, *37*, 283–285.

Hu, T., Sinnott-Armstrong, N. A., Kiralis, J. W., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2011). Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, *12*, 1–13.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians*, *61*, 69–90.

Jeong, D., Park, S., Kim, H., Kim, C.-J., Ahn, S. T., Bae, S. B., Kim, H. J., Kim, T. H., Im, J., Lee, M. S., Kwon, H. Y., & Baek, M. J. (2016). Rhoa is associated with invasion and poor prognosis in colorectal cancer. *International Journal of Oncology*, *48*(2), 714–722. https://doi.org/10.3892/ijo.2015.3281

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., & Lempicki, R. A. (2012). DAVID-WS: A stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, *28*, 1805–1806.

Kafaie, S., Chen, Y., & Hu, T. (2019). A network approach to prioritizing susceptibility genes for genome-wide association studies. *Genetic Epidemiology*, 1–15. https://doi.org/10.1002/gepi.22198

Loos, R. J., & Yeo, G. S. (2014). The bigger picture of FTO - the first GWAS-identified obesity gene. *Nature Reviews Endocrinology*, *10*, 51.

Ma, Y., Zhang, P., Wang, F., Liu, W., Yang, J., & Qin, H. (2012). An integrated proteomics and metabolomics approach for defining oncofetal biomarkers in the colorectal cancer. *Annals of Surgery*, *255*, 720–730.

Manduchi, E., Chesi, A., Hall, M. A., Grant, S. F. A., & Moore, J. H. (2018). Leveraging putative enhancer-promoter interactions to investigate two-way epistasis in Type 2 Diabetes GWAS. In: Lussier, Y. A., Berghout, J., Vitali, F., Ramo, K. S., Kann, M., & Moore, Jason H. (eds), *Pacific Symposium on Biocomputing* (Vol. 23, p. 548–558). Stanford University.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., ... Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747.

Martins, M., Rosa, A., Guedes, L. C., Fonseca, B. V., Gotovac, K., Violante, S., Mestre, T., Coelho, M., Rosa, M. M., Martin, E. R., Vance, J. M., Outeiro, T. F., Wang, L., & Oliveira, S. A. (2011). Convergence of miRNA expression profiling, α -synuclein interacton and GWAS in Parkinson's disease. *PLOS ONE*, *6*(10), e25443.

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, *56*, 73–82.

Murk, W., & DeWan, A. T. (2016). Exhaustive genome-wide search for SNP-SNP interactions across 10 human diseases. *G3 (Bethesda, Md.)*, *6*(7), 2043–2050.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*, 559–575.

Raghavan, N., & Tosto, G. (2017). Genetics of Alzheimer's disease: The importance of polygenic and epistatic components. *Current Neurology and Neuroscience Reports*, *17*, 78.

Rodrigues, P., Macaya, I., Bazzocco, S., Mazzolini, R., Andretta, E., Dopeso, H., Mateo-Lozano, S., Bilić, J., Cartón-García, F., Nieto, R., Suárez-López, L., Afonso, E., Landolfi, S., Hernandez-Losa, J., Kobayashi, K., Ramón y Cajal, S., Tabernero, J., Tebbutt, N. C., Mariadason, J. M., ... Arango, D. (2014). Rhoa inactivation enhances wnt signalling and promotes colorectal cancer. *Nature Communications*, *5*, 5458.

Schubert, B., Maddamsetti, R., Nyman, J., Farhat, M. R., & Marks, D. S. (2019). Genome-wide discovery of epistatic loci affecting antibiotic resistance in Neisseria gonorrhoeae using evolutionary couplings. *Nature Microbiology*, *4*, 328–338.

Schüpbach, T., Xenarios, I., Bergmann, S., & Kapur, K. (2010). FastEpistasis: A high performance computing solution for quantitative trait epistasis. *Bioinformatics*, *26*, 1468–1469.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*, 2498–2504.

Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, *68*(1), 7–30.

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H., & Sun, Y. V. (2009). Machine learning in genome-wide association studies. *Genetic Epidemiology*, *33*, S51–S57.

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., Barclay, E., Lubbe, S., Martin, L., Sellick, G., Jaeger, E., Hubner, R., Wild, R., Rowan, A., Fielding, S., ... Houlston, R. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24. 21. *Nature Genetics*, *39*, 984.

Tu, R., Kang, W., Yang, X., Zhang, Q., Xie, X., Liu, W., Zhang, J., Zhang, X.-D., Wang, H., & Du, R.-L. (2018). USP49 participates in the DNA damage response by forming a positive feedback loop with p53. *Cell Death & Disease*, *9*(5), 553. https://doi.org/10.1038/s41419-018-0475-3

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., & Yu, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, *87*, 325–340.

Wang, Y., Liu, G., Feng, M., & Wong, L. (2011). An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics*, *27*(21), 2936–2943.

Xie, Y., Chen, L., Lv, X., Hou, G., Wang, Y., Jiang, C., Zhu, H., Xu, N., Wu, L., Lou, X., & Liu, S. (2016). The levels of serine proteases in colon tissue interstitial fluid and serum serve as an indicator of colorectal cancer progression. *Oncotarget*, *7*(22), 32592–32606.

Xie, W., Rimm, D. L., Lin, Y., Shih, W. J., & Reiss, M. (2003). Loss of Smad signaling in human colorectal cancer is associated with advanced disease and poor prognosis. *Cancer Journal (Sudbury, Mass.)*, *9*(4), 302–312.

Zeng, Q., Phukan, S., Xu, Y., Sadim, M., Rosman, D. S., Pennison, M., Liao, J., Yang, G.-Y., Huang, C.-C., Valle, L., DiCristofano, A., de la Chapelle, A., & Pasche, B. (2009). Tgfbr1 haploinsufficiency is a potent modifier of colorectal cancer development. *Cancer Research*, *69*(2), 678–686.

Zhang, X., Huang, S., Zou, F., & Wang, W. (2010). TEAM: Efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, *26*, i217–i227.

Zhou, R., Huang, Y., Cheng, B., Wang, Y., & Xiong, B. (2018, 03). Tgfbr1*6a is a potential modifier of migration

and invasion in colorectal cancer cells. *Oncology Letters*, *15*(3), 3971–3976.

Zhou, G., Yang, J., & Song, P. (2019). Correlation of ERK/MAPK signaling pathway with proliferation and apoptosis of colon cancer cells. *Oncology Letters*, *17*(2), 2266–2270.

Zhu, Y.-P., Liu, Z., Fu, Z.-X., & Li, D.-C. (2017, 03). SMAD3 mutant mice develop colon cancer with overexpression of cox-2. *Oncology Letters*, *13*(3), 1535–1538.

How to cite this article: kafaie S, Xu L, Hu T. Statistical methods with exhaustive search in the identification of gene–gene interactions for coloretctal cancer. *Genetic Epidemiology*. 2021;45:222–234. https://doi.org/10.1002/gepi.22372