Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis

Julie E. Goodman¹, Leah E. Mechanic¹, Brian T. Luke², Stefan Ambs¹, Stephen Chanock³, and Curtis C. Harris¹,*
¹ Laboratory of Human Carcinogenesis, National Cancer Institute, NIH, Bethesda, MD, USA
² Frederick Cancer Research Facility, National Cancer Institute, NIH, Frederick, MD, USA
³ Pediatric Oncology Branch, National Cancer Institute, Advanced Technology Center, Gaithersburg, MD, USA

Abstract
Several single nucleotide polymorphisms (SNPs) in genes derived from distinct pathways are associated with colon cancer risk; however, few studies have examined SNP-SNP interactions concurrently. We explored the association between colon cancer and 94 SNPs, using a novel approach, polymorphism interaction analysis (PIA). We developed PIA to examine all possible SNP combinations, based on the 94 SNPs studied in 216 male colon cancer cases and 255 male controls, employing 2 separate functions that cross-validate and minimize false-positive results in the evaluation of SNP combinations to predict colon cancer risk. PIA identified previously described null polymorphisms in glutathione-S-transferase T1 (GSTT1) as the best predictor of colon cancer among the studied SNPs, and also identified novel polymorphisms in the inflammation and hormone metabolism pathways that singly or jointly predict cancer risk. PIA identified SNPs that may interact with the GSTT1 polymorphism, including coding polymorphisms in TP53 (Arg72Pro in p53) and CASP8 (Asp302His in caspase 8), which may modify the association between this polymorphism and colon cancer. This was confirmed by logistic regression, as the GSTT1 null polymorphism in combination with either the TP53 or the CASP8 polymorphism significantly alter colon cancer risk (pinteraction < 0.02 for both). GSTT1 prevents DNA damage by detoxifying mutagenic compounds, while the p53 protein facilitates repair of DNA damage and induces apoptosis, and caspase 8 is activated in p53-mediated apoptosis. Our results suggest that PIA is a valid method for suggesting SNP-SNP interactions that may be validated in future studies, using more traditional statistical methods on different datasets (Supplementary material can be found on the International Journal of Cancer website at http://www.interscience.wiley.com/jpages/0020-7136/suppmat).

Keywords
polymorphism interaction analysis; single nucleotide polymorphism; colon cancer

Glutathione-S-transferase T1 (GSTT1) prevents DNA damage by detoxifying mutagenic compounds. The null polymorphism in this GST gene has been hypothesized to increase cancer risk, because decreased enzyme activity leads to impaired metabolic elimination of...
carcinogens.\(^1\) Besides \textit{GSTT1}, additional single nucleotide polymorphisms (SNPs) have been associated with colon cancer risk, such as \textit{TNF-\(\alpha\)}, \textit{NAT2}, and \textit{HRAS1}.\(^2\) de Jong \textit{et al.}\(^2\) reviewed published studies between 1980 and 2001, and performed pooled analyses for 30 SNPs in 20 genes in colon cancer studies. Overall, the analyses found modest effects for colon cancer risk for a few of the SNPs. For example, \textit{GSTT1} null was associated with an increased colon cancer risk in the pooled analysis, but the finding was not significant in many of the individual studies. In 3 of 10 studies with significant results, the odds ratios (ORs) were reported to vary greatly from 1.37 (95% confidence interval (CI): 1.17–1.60) to 4.49 (95% CI: 2.42–8.34).\(^3\)–\(^5\)

Inconsistencies among these SNP studies could be due to each individual SNP altering the function of only 1 gene of the many that are involved in carcinogenesis. The biological events associated with cancer risk that are modestly affected by a single SNP may be more greatly affected by a SNP in combination with additional SNPs or biological or environmental factors. In addition, a SNP, which has no obvious effect on colon cancer risk, may play a significant role in the presence of other SNPs/factors. Hence, these studies could have benefited from examining interactions between polymorphisms in distinct genes, and by using a method that could examine high order SNP-SNP interactions and disease risk.

There are currently several algorithms designed to examine complex SNP-SNP interactions and disease risk.\(^6\) These methods include multifactor dimensionality reduction (MDR),\(^6\)–\(^9\) which examines all possible SNP combinations from a set of given SNPs and chooses the combination that best predicts risk by minimizing the classification error of cases and controls. We created polymorphism interaction analysis (PIA), which, similar to MDR, examines all possible SNP combinations from a given set of SNPs. PIA differs as it uses 2 unique scoring functions, the Gini Index and \% wrong, to identify the SNP-SNP combinations that are most likely associated with disease risk.

We have used PIA to identify SNP-SNP interactions among 94 SNPs in 63 genes drawn from pathways of angiogenesis, apoptosis, DNA repair, inflammation, hormone metabolism and general metabolism. While certainly not an exhaustive list of all existing common SNPs, each was chosen because of a high prior, such as in or affecting a cancer-related gene of demonstrated involvement in the aforementioned pathways, previous functional analysis or a previous association study related to cancer risk. PIA is not intended to be a stand-alone method, but rather to narrow down the number of potential SNPs and SNP combinations to be examined using more traditional statistical methods in other datasets. This method identified several SNPs, which in combination may predict colon cancer risk, such as SNPs that interacted with the \textit{GSTT1} null polymorphism, including coding polymorphisms in \textit{TP53} (Arg72Pro in p53) and \textit{CASP8} (Asp302His in caspase 8). We validated the results of PIA internally and by crossvalidation with logistic regression and \(\chi^2\) testing. The data provided shows that PIA is a robust tool for examining SNP combinations.

**Material and methods**

**Study population**

Subjects consist of males that participated in a larger colon cancer study, which was described in detail previously.\(^10\) Briefly, cases (\(n = 216\)) were recruited between 1992 and 2003, and controls (\(n = 255\)) were recruited between 1998 and 2003. All subjects resided in the greater Baltimore area at the time of recruitment. Cases and hospital controls were recruited from Baltimore area hospitals, while population controls were identified from Maryland Vehicle Administration records and contacted by mail, and then by telephone. Inclusion criteria included living in the greater Baltimore area, being Caucasian or African American, and being born in the United States. Subjects were excluded if, according to self report, they had other cancers, or were infected with HIV, HBV or HCV, or were IV drug users, were
institutionalized, were diagnosed over 6 months before the study, or had a disability. The ethnic group was classified according to self report. Previous analyses demonstrated that cases recruited before 1998 were similar to those recruited after, suggesting that the difference in time period for case and control selection does not influence results reported here. All subjects signed informed consent forms and were administered an epidemiology questionnaire. This protocol was approved by the NCI IRB and the IRBs of all participating institutions.

**DNA isolation and genotyping**

Blood was collected when possible from both cases and controls for DNA extraction, using the Qiagen FlexiGene DNA Kit (Qiagen, Valencia, CA); however, in some cases when blood was not available, colon tissue was used to obtain DNA (for cases) and DNA was isolated using the DNeasy Tissue Kit (Qiagen). It is very unlikely that a mutation would have occurred at the exact base of one of the SNPs of interest, causing an altered genotype; therefore, the potential artifact in archived specimen genotyping is minimal. All genotyping was performed at the National Cancer Institute Core Genotyping Facility (NCI CGF), using validated assays that can be found on their website, http://snp500cancer.nci.nih.gov. Genotyping for the GSTT1-02 polymorphism was performed using a real-time, PCR-based approach developed at NCI CGF, to distinguish between individuals with 1 or 2 copies of the GSTT1-02 locus. This is a method similar to an approach described previously, but differs from traditional methods that detect the presence versus absence of these alleles. Detailed procedures are available at the NCI CGF web site.

We focused on 63 genes in pathways known to have a role in colon carcinogenesis. The 94 SNPs in these genes were chosen based on the following criteria: (i) the allele variant gene is potentially functionally important, based on previous publications or because the allele variant results in a change of the amino acid sequence of the protein; (ii) the polymorphism will likely affect protein expression/stability/activity or mRNA splicing/stability; (iii) an association between the genotype and cancer risk was previously shown; and (iv) the variant allele is common (> 5%). All SNPs are described in Supplemental Table I (Supplementary material can be found on the International Journal of Cancer web-site at http://www.interscience.wiley.com/jpages/0020-7136/suppmat). Of the SNPs, 85% had 5% or fewer samples that failed to genotype. The remainder of the SNPs had between 6–9% samples that failed to genotype. Samples that failed to genotype were scored as missing. Ten percent of all samples were genotyped twice for quality control, and concordance was 100% for 84 (90%) of the SNPs. Concordance was 97% for CYP19A1-04, GSTP1-01, XRCC3-01, DIO1-04, PLAB-02 and EDN1-02; 94% for ILB-03, IL4R-02 and CASP9-01; and 91% for MTHFR-02.

**Statistical analyses**

Logistic regression and \( \chi^2 \) analyses were performed using Stata 7.0 (Stata Corp, College Station, TX). Age and body mass index (BMI) were compared using the Student's t-test. ORs, 95% CIs and \( p_{\text{trend}} \) and \( p_{\text{interaction}} \) values were calculated using unconditional logistic regression. Departures from Hardy–Weinberg equilibrium were evaluated using \( \chi^2 \) tests. With the exception of MAOA-01, all SNPs were in Hardy–Weinberg equilibrium in at least 1 ethnic control group, while 86% of SNPs examined were in Hardy–Weinberg equilibrium in both Caucasian and African-American controls. Because cases were recruited between 1992 and 2003 and controls were recruited between 1998 and 2003, certain analyses were performed using the subset of the cases recruited from 1998 to 2003, which yielded similar results to those using the whole case group (data not shown). In addition, results were similar for analyses using hospital and population controls separately, and therefore, they are combined for all analyses presented.
Polymorphism interaction analysis

We developed PIA to examine all possible N SNP combinations to find high order SNP-SNP interactions that best predict disease risk. We are currently in the process of making PIA available on the internet and will publish a manuscript with more detailed methods (manuscript, in preparation). Here, we give a general description of PIA.

PIA uses 2 separate scoring functions, which each independently determines how well a particular SNP combination predicts risk by differentiating cases and controls. PIA then stores the 50 best combinations ordered from lowest to highest score for each (the lower the score, the better at predicting risk). These 50 SNP combinations are also examined to determine whether a particular SNP appeared regularly, either alone or in combination with other SNPs. Next, results from both scoring functions are compared to determine which SNP-SNP interactions appeared using both approaches, to allow for crossvalidation and to limit the appearance of false-positive associations.

There were several SNPs for which genotyping data were missing for cases and controls. If the number missing varies greatly between cases and controls, results could be affected by differential missing values, i.e., a SNP or SNP combination may appear predictive of case or control status when in reality this SNP or SNP combination was scored highly because of differences in the percent of missing genotypes in cases compared with controls. Therefore, for both scoring functions, if for a certain SNP or set of SNPs, the case/control or control/case ratio was greater than 1.2, this SNP or SNP combination was not analyzed to prevent missing values from leading to false-positive results. The value 1.2 was chosen as a conservative threshold.

The first score metric, Gini, uses the Gini Index, which is utilized in all classification and regression tree decision trees,13 to determine which feature is used to separate a given node. In general, the sample set is separated in C nodes or cells based on the combination of SNPs (or genotypes) examined. For example, if examining a set of 2 SNPs, a subject can be homozygous wild-type, heterozygous or homozygous variant at each of the 2 sites, leading to 9 combinations or a total of 9 cells. The distribution of cases and controls across these cells determines the score of the SNP combination. For cell c, the Gini scoring function is determined using the expression:

$$\text{GINI}(c) = 1.0 - \left(\frac{n_{ca}(c)^2 + n_{co}(c)^2}{n_t(c)^2}\right)$$

In this expression, $n_{ca}(c)$ and $n_{co}(c)$ are the number of cases and controls in each cell, respectively, and $n_t(c)$ is the total number of subjects in each cell. Using GINI(c), the score of the overall split into C cells is given by the GINI_split formula:

$$\text{GINI}_{\text{split}} = \sum_{c=1}^{C} \left[\left(\frac{n_t(c)}{n_T}\right)\text{GINI}(c)\right]$$

Here $n_T$ is the total number of samples, whose genotypes are identified in all selected SNPs. Therefore, the GINI_split is a weighted average of GINI(c) across all cells for each SNP combination and can range from 0 to 1. The lower the score, the better the model is at predicting risk.

The second scoring function, % wrong, determines the percentage of wrong assignments, i.e., the incorrect designation of case or control status, for the samples in the dataset. This method bears similarities to MDR,6–9 with several novel differences, including the distribution of the data into testing and training sets, how PIA deals with missing data and what PIA does...
with genotypes for only 1 subject, as discussed later. In this application, the cases and controls are scrambled and divided into 10 sets by dividing the sample by 10. However, if a sample set is not divisible by 10, the subjects are divided in such a way that there are 10 sets, but not necessarily all with the exact same number of subjects. This method ensures that the entire sample set is used and does not need to be exactly divisible by 10. The 10 randomized units are used to create 10 distinct training and testing sets. The training sets are first used to select SNP combinations that best predict risk, and these combinations are next examined in the testing sets, as discussed later. The first testing set uses the first approximate 10% of the subjects from the scrambled dataset, and the corresponding training set is the remaining approximate 90% of the subjects. The second testing set is the second approximate 10% of the subjects, and so on. The dataset is used to construct 10 training/testing set combinations such that all training/testing sets have approximately the same fraction of cases and controls. For each set of SNPs, the samples in the training set are placed in one of the C cells. The number of misclassified samples is initially set to zero in each cell examined. Once again, \( n_{\text{cad}}(c) \) and \( n_{\text{co}}(c) \) are the number of cases and controls in the training set in cell \( c \), respectively. If \( n_{\text{cad}}(c) \) is less than or equal to \( n_{\text{co}}(c) \), then the number of misclassified is \( n_{\text{cad}}(c) \), otherwise, the number of misclassified is \( n_{\text{co}}(c) \). If a cell contains an equal number of cases and controls from the training set, they are all 50% misclassified. In addition, this scoring function is designed to penalize singletons, which are cells containing only a single sample. If cell \( c \) contains a single sample, the number of misclassified in the training set is increased by one. The total number misclassified is called the classification error.

Once the number of misclassified training samples is determined across all cells, each of the samples in the test set is examined. If a sample in the test set is incorrectly classified, it contributes to the prediction error. The sample is placed into the appropriate cell, and the sample is correctly classified only if the number of training samples of its type (case or control) in this cell is greater than the number of training samples of the other type. Otherwise, the prediction error is increased by one. Any testing sample that is placed in a cell containing an equal number of cases and controls is considered misclassified.

The percent wrong is the sum of the classification error plus the prediction error divided by the total number of samples in the training and testing sets, and like the GINI split, can range from 0 to 1. This was calculated for each of the 10 training/testing set combinations. The samples were then rescrambled and the entire process was repeated for a total of 10 cycles. The final percent wrong calculated was the average of the percent wrong values calculated for each of the 100 training/testing set combinations. Using the percent wrong function, PIA identifies those SNP combinations that produce the smallest percent wrong. The score for this scoring function is the percent of subjects misclassified. Therefore, a lower score means fewer people are misclassified, and hence, a better model.

**Results**

Cases and controls were similar with respect to age, BMI and ethnicity (\( p > 0.2 \) for all). The average age of cases was 65.6 ± 11.3 years and that of controls was 66.4 ± 9.7 years. The average BMI for cases was 27.4 ± 5.7 kg/m\(^2\) and 28.0 ± 5.1 kg/m\(^2\) for controls. African Americans made up 35.3% of cases and 37.5% of controls, while all other subjects were Caucasians.

The associations between colon cancer risk and 94 individual SNPs in 63 genes (listed in Supplemental Table I) were initially examined by unconditional logistic regression. The associations that approached (\( p < 0.1 \)) or achieved (\( p < 0.05 \)) statistical significance are shown in Table I. Initially, several individual SNPs were associated with colon cancer risk. We
observed increased risk associated with *CYP19A1-14, MTRR-01, MTHFR-02, IL1B-01, IL4-01, TP53-01* and *GSTT1-02*. Results were similar for crude analyses and adjusted for age and ethnicity (data not shown).

After examining the main effects of each SNP, the goal was to examine potential interactions between these SNPs. Moreover, it was of interest to identify interactions that predict colon cancer risk, even in the absence of main effects. However, with such a large number of SNPs, logistic regression was not practical or powerful enough to analyze all high order SNP-SNP interactions (from the 94 SNPs) and colon cancer risk. Therefore, PIA was used for these 94 SNPs using 2 scoring functions, Gini and percent wrong, to determine the SNP combinations that best predicted risk. The lower the score of a scoring function, the better the model was at predicting risk. Ethnicity was also included in the analysis, because it was considered a potential confounder. PIA was used to examine the association of each individual SNP (1st order SNP-SNP interaction), and all possible combinations of 2 SNPs (2nd order), 3 SNPs (3rd order) and 4 SNPs (4th order) with colon cancer, to determine which combination of SNPs were best at classifying colon cancer cases and controls. In other words, if PIA was used for a 4 SNP combination (4th order), it examined all possible combinations of 4 SNPs and determined which combinations were the best at determining colon cancer risk.

For each of the 1st through 4th order SNP-SNP interaction models, PIA ranked each combination of SNPs according to the quality of the models. The top models, or those with the lowest score, are shown in Table II. These models were also validated with $\chi^2$ analyses, which showed that the associations were all statistically significant ($p < 0.03$ for all). Because PIA does not provide a magnitude of association, the risk of colon cancer associated with high-risk versus low-risk genotypes were calculated crudely by logistic regression (Table II). This was done by classifying each genotype as high risk, if more cases than controls had it, and low risk, if more controls than cases had it. As expected, all high-risk groups led to a small, but significant, increased risk of colon cancer. This risk increased as the number of SNPs in the model increased. These values were similar when analyses were adjusted for age and ethnicity (data not shown).

It is more likely that multiple pathways contribute to colon carcinogenesis; therefore we were interested in examining the top 50 models for predicting cancer risk. Interestingly, the most robust models for 1st through 4th order interactions shown in Table II were only incrementally better, but not significantly different, from other models tested (Table III, only the top 20 models are shown). Overall, the range in scores was between 0.36 and 0.48 for the Gini function and 0.32 and 0.42 for the percent wrong function. It is not surprising that as more SNPs are added to the model, the results yielded lower scores, because of the inclusion of additional terms in the model explaining more variation.

The PIA model is robust in our analysis of the top 20 models for the 1st order SNPs, because our results are comparable to those obtained using logistic regression (Tables I and III). When comparing the top 20 models for a single SNP to the data in Table I, 14/20 SNPs were identified by the Gini scoring function, and 10/20 of SNPs were identified by the % wrong scoring function. In this regard, both performed comparably as classifiers of case control status, using PIA and logistic regression. Moreover, the results obtained using the Gini scoring and the percent wrong scoring functions were similar. For the 1st, 2nd and 4th order SNP-SNP interaction models, 13, 11 and 9 out of 20 models, respectively, were identical, while additional couplings of SNPs were similar in even more models (Table III, bold).

The SNPs that were most frequently selected in each of the top 50 models represent a set with a high probability for interaction, and perhaps main effects to be tested separately. Each SNP
was ranked according to the frequency by which it appeared in each analysis for the 1st to 4th order SNP combinations, using both the Gini and % wrong functions and the rank of these models. This was then averaged among the 8 models to produce a combined ranking for each individual SNP (Table IV). There were many consistencies between both scoring functions and also among the 1st to 4th order SNP-SNP interaction models—many of the SNPs most frequently observed in the top 50 models were similar, using separate scoring functions. For example, GSTT1-02, IL4-01, TP53-01, SOD2-01, IL10-02 and MTRR-01 ranked among the top 10 for most models. Of the top 10 most frequently appearing individual SNPs, as determined by ranking in Table IV, 7 (IL4-01, OGG1-04, TP53-01, GSTT1-02, SOD2-02, IL1B-01 and IL10-02) were identified by logistic regression (Table I). However, logistic regression did not detect main effects for WRN-03, GSTP1-01 or CYP19A1-06, nor did it identify the numerous SNPs that interact with the top SNPs using PIA.

In our PIA analysis of the colon cancer dataset, there were a series of SNPs that appear to interact with GSTT1-02, namely, TP53-01, IL4-03 and CASP8-03 in the 2nd order models by both scoring functions (Tables II and III). These interactions were later confirmed using logistic regression ($p_{interaction} < 0.02$ for all) and by $\chi^2$ analyses ($p < 0.04$ for all). In the 3rd order models, GSTT1-02 interacted with ESR1-03 and GPX1-03, or with TP53-01 and IL10-02 ($p < 0.02$ for both by $\chi^2$ analyses). Three-way and 4-way interactions could not be validated using logistic regression modeling, partly due to inadequate power. However, the ability of all of these combinations of SNPs to differentiate between cases and controls was validated by $\chi^2$ analyses ($p < 0.05$ for all).

**Discussion**

Several individual SNPs have been associated with colon cancer risk,² thus we were interested in identifying which combinations of SNPs among 94 SNPs of interest in 63 genes were associated with colon cancer risk. These SNPs were selected because they may play a role in colon cancer etiology, and it was previously unknown whether a combination of several of these SNPs increased risk to a greater extent than a particular SNP alone. We created PIA with 2 scoring functions to determine which SNP combinations were associated with risk. Both scoring functions have a high level of agreement and are effective in determining SNP-SNP interactions, including some involving SNPs with no main effects that were associated with colon cancer risk.

PIA identified GSTT1-02, which generates a null GSTT1 allele, in several of the top models (those with the lowest scoring function). de Jong et al.² discussed 10 studies that examined the GSTT1-null allele and colon cancer risk, and found it was inconsistently associated with an increased colon cancer risk. Perhaps the lack of reproducible association with colon cancer could be related to interactions with additional SNPs, including those identified by PIA. These include both TP53-01, which results in an Arg72Pro change in the p53 protein, and CASP8-03, which is an Asp302His in caspase 8. The Arg72Pro polymorphism of TP53 was shown to modulate the p53-dependent apoptotic pathway through mitochondrial targeting in a transcription-independent manner.¹⁴ It has been shown that caspase 8 plays an essential role in the transcriptional-independent pathway of p53-mediated apoptosis.¹⁵,¹⁶ Therefore, alterations in these pathways could lead to increased colon cancer risk by deregulating apoptosis. In addition, 2 studies examined interactions between the GSTT1 polymorphism and TP53 mutations, in particular, cancers.¹⁷,¹⁸ These studies found that individuals with a GSTT1-null genotype were more likely to have TP53 mutations in esophageal adenomas and breast cancer,¹⁷ perhaps through inefficient detoxification of DNA damaging agents. A hypothesis can be generated from these PIA results and the available literature that SNPs in GSTT1 in combination with either CASP8 or TP53 together affect cancer risk more than each
SNP alone. A GSTT1-null individual may have more TP53 mutations, which, in conjunction with CASP8 or TP53 polymorphisms, leads to a decreased efficiency of p53-mediated apoptosis resulting in a reduced ability to remove damaged cells and a higher colon cancer risk. In addition to GSTT1-02, PIA identified several other individual SNPs that interact with other SNPs in different models, suggesting the participation of individual SNPs in several distinct pathways. Moreover, several different combinations of SNPs in these genes were shown by PIA to be predictive of colon cancer risk. Taken together, these results are consistent with the notion that it is more likely that the carcinogenic process differs among individuals, and that colon carcinogenesis occurs through a variety of interconnected pathways. Furthermore, the same pathway may be altered by several SNPs; therefore, distinct SNP sets might actually represent the same biological outcome because they are affecting the same pathway.

While the focus of PIA is to identify significant interactions in colon carcinogenesis, several individual SNPs appeared with high frequency in the 1st through 4th order models, using both scoring functions. This result suggests that these SNPs are important individually, as well as in combination with other SNPs. Moreover, these results could also indicate that these SNPs act as molecular nodes or junctions between multiple pathways of colon carcinogenesis. Therefore, PIA serves as both a powerful tool for identifying interactions and for identifying those single SNPs, which function in several biological pathways.

One of the major strengths of PIA, as shown by the analysis presented here, is that a crossvalidation process based on repetitive modeling of the data can verify many of the associations. Moreover, analyses by 2 parallel scoring functions provide internal crossvalidation. For instance, many of the top 50 models (1st to 4th order combinations) were similar using these 2 functions, as were the most frequently appearing single SNP results. In addition to internal crossvalidation by PIA, many of the single SNPs identified in the top 20 models using both functions of PIA were observed by logistic regression modeling to be associated with colon cancer risk. Furthermore, using logistic regression, p-interaction for the 2 SNP combinations identified by PIA were determined to be statistically significant. Higher order interactions, or the combinations of SNPs, were confirmed to be predictive of case or control status, using $\chi^2$ tests of association. Finally, for the top models identified by PIA, logistic regression modeling produced ORs consistent with the observed association with colon cancer risk (Table II).

In this study, the ORs in Table II are based on comparing all genotypes with more cases than controls against all genotypes with more controls than cases (2 categories). In contrast, in many epidemiology studies examining SNPs and cancer risk, SNP combinations are often stratified in several categories. In these studies, the highest risk genotype combination is often compared with the lowest risk genotype combination to estimate an OR, which produces the largest possible OR. In our study, however, because all risk categories are grouped into only 2 categories, this necessarily diminishes the magnitude of the OR as opposed to comparing extreme values. Incorporating larger combinations of SNPs, such as 5 or 6, would only increase the estimated OR slightly, because we would still be comparing 2 broad categories of SNPs instead of extremes. Moreover, PIA does not have the power to examine these high order combinations accurately, because as the number of SNPs in a combination (or the model order) increases, the number of cases and controls with each genotype combination decreases, increasing the likelihood of many genotype combinations with zero individuals. Given this limitation, much remains to be learned from the 1st through 4th order combinations of SNPs, and both examining this number of SNPs in combination, and simultaneously investigating 94 SNPs is advantageous to a traditional logistic regression approach requiring a one by one examination of interaction between SNPs.
PIA is a useful tool to detect multiple interactions between polymorphisms. For instance, PIA detected interactions among several SNPs in absence of main effects of these SNPs, including CASP8-03 with GSTT1-02 as mentioned earlier. Using logistic regression modeling, these interactions may have been overlooked. In addition, examining the top 50 SNP combinations has many advantages over examining the top combination alone. Since the scores are very similar for the top 50 models (Table III), in addition to the top model, several of the successive models quite likely describe SNP combinations that contribute to cancer risk to a similar degree as the top model.

Another strength of PIA is in examining the 1st to 4th order SNP combinations. This is illustrated by IL1B-01 and IL1B-03. Both scoring functions identified IL1B-01 and IL1B-03 as high-ranking 2nd order models. However, the combination of IL1B-01 and IL1B-03 appeared in 45 of the top 50 Gini 3rd order SNP models, but did not appear in the 3rd order percent wrong SNP models or in any of the top 50 fourth order SNP models. The reason it appears in the 3rd order Gini Cost function in each of the top 50 models is that the Gini algorithm determined that the combination of both of these SNPs was strong enough that no combination of 3 SNPs better predicted risk than this 2 SNP combination. The 4th order model, however, showed that combinations of 4 SNPs were stronger predictors of risk than the IL1B-01 and IL1B-03 together, as these 2 SNPs were not in any of the top 50 fourth order combinations, using the Gini method. Because the percent wrong scoring function uses a different algorithm than the Gini method, it did not find that the IL1B-01 and IL1B-03 combination was stronger than combinations of 3 SNPs, and hence, the IL1B-01 and IL1B-03 combination was not in any of the 3rd order SNP models. These results suggest that the effects of these polymorphisms in combination may be diluted in the presence of additional polymorphisms and stress the advantage of examining the 1st through 4th order models with 2 methods, as SNP combinations might not have been realized examining higher order models alone or with 1 method.

When examining such a large number of SNPs and SNP combinations, there is always a possibility that a SNP combination is selected due to chance or problems in study design or analysis. Using 2 scoring functions for verification reduces the effect of selecting a SNP combination, because this combination accidentally produces the largest difference between known cases and controls. $\chi^2$ analyses or logistic regression modeling confirmed many of the associations observed using PIA. The formal possibility of chance associations may only be addressed by confirmation in other studies of colon cancer.

PIA looks for subcombinations that consistently appear in the top models, it is not substantially affected by sample size. In this regard, thus it is well suited as a tool for exploratory analyses. In comparison, $\chi^2$ analysis shows that a particular combination produces a non-random distribution of cases and controls, and this should remain valid as the number of samples increases; assuming that the new samples are taken from the same population used in the current study. While the order of the top models might change slightly because the scores are similar, the same models should appear.

Our dataset was ideal for PIA in which cases and controls were well matched for potential confounders (age and ethnicity) and in which there were a similar number of cases and controls. However, a limitation of PIA is that it cannot account for confounding variables. To determine whether a confounder modulates an association between SNPs and disease, selected SNP combinations can be analyzed by generating a dichotomous variable of high- and low-risk genotypes (see Table II), and modeled using logistic regression. Another approach is to include potential confounders as a variable in PIA, as was done for ethnicity in this study. While SNPs have been the focus of this study, it is possible to test whether any binary or trinary factor significantly interacts with each other or with a SNP. A third approach for dealing with a known
confounder is to stratify data accordingly and perform analyses for each stratum separately. This approach is advisable, because otherwise the confounder will not necessarily be selected in the top models. In our study, we limited our dataset to men because gender is a potential confounder of colon cancer risk.

There are currently several algorithms designed to examine SNP interactions and disease risk, including MDR, which bears similarity to PIA. For example, similar to the % wrong scoring function, MDR uses a testing set with the first 10% of the subjects from a scrambled dataset and a corresponding training set with the remaining 90% of the subjects, and does this for all 10% units of the data and then repeats this 10 times. MDR also examines every possible SNP combination and chooses the one with the minimum misclassification. However, in its current form, MDR requires that the total number of subjects be divisible by the number of test sets, and data may have to be dropped to meet this requirement. In addition, MDR is very sensitive to missing values, particularly if the number of missing values differs in cases and controls. If the number of missing values, but not the genotypes, differs between cases and controls, MDR will identify a SNP set as a good predictor of risk, although in reality it is not. PIA was designed to be less sensitive to missing values, as it eliminates a SNP set if the difference in missing values in cases and controls exceeds a certain threshold (selected in this study to be a case/control or control/case ratio greater than 1.2). A possible disadvantage to this approach is that with an increase in specificity, there is a decrease in sensitivity. An interesting SNP set may not be identified because it was dropped from the testing set due to missing values. Another advantage of PIA in comparison with MDR is that PIA uses 2 separate scoring functions to allow for crossvalidation, and PIA examines the top 50 models (versus the top single model) to increase sensitivity to detect key interactions.

By definition, it is highly unlikely that a single SNP will efficiently discriminate between cases and controls in a colon cancer study; however, it is plausible that a set of SNPs derived from genetic pathways critical in colon carcinogenesis could contribute to risk. While risks for any single SNP are relatively small, it is necessary to develop robust analytical tools to identify and validate the constituents of the set of informative SNPs or group of SNPs. It should be noted that because SNPs have small effects, it suggests that not only do they play a role in carcinogenesis, but also in the genes and biological pathways in which they reside that are crucial to this process. Knowledge of major SNP-SNP interactions provides insight into the relationship of complex pathways and also highlights key genes that could be targets for therapy or preventive strategies.

Finally, PIA identified several SNPs which have been previously shown to be associated with colon cancer risk, such as GSTT1 and MTHFR polymorphisms. This validates PIA as an analysis tool, because these findings have been reported previously. In addition, PIA identified several novel interactions that have not been analyzed previously. For example, PIA identified interactions among SNPs in genes encoding interleukins and in genes encoding proteins involved with hormone metabolism. This novel finding should be followed up in future studies.

In summary, we have used PIA to identify SNP-SNP interactions that may be critical in colon cancer risk. These included SNPs that did not appear to singularly contribute to colon cancer risk. PIA is not designed to stand on its own, but rather to narrow down a vast list of potentially key SNPs and SNP combinations to a more manageable size that can be explored using more traditional methods in other datasets. While we have provided internal validation of a subset of the tested SNP combinations in our study population, these interactions require crossvalidation in other populations and should form the rationale for investigating the
functional basis of putative causal SNPs in the laboratory. Lastly, we have shown that PIA is a robust method for determining SNP-SNP interactions in association studies.

Supplementary Material

Supplementary Table I.

Acknowledgements

We thank Dorothea Dudek for editorial assistance, Karen MacPherson for bibliographical assistance, and our collaborators at UMD and JHU.

References

## TABLE I

SNPs WITH SIGNIFICANT \((p < 0.05)\) OR NEAR-SIGNIFICANT \((p < 0.1)\) ASSOCIATIONS WITH COLON CANCER

<table>
<thead>
<tr>
<th>SNP</th>
<th>Wild type allele (aa)</th>
<th>Variant allele (aa)</th>
<th>Wild type/wild type OR (95% CI)</th>
<th>Wild type/variant OR (95% CI)</th>
<th>Variant/variant OR (95% CI)</th>
<th>(p_{\text{trend}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEK1-02</td>
<td>G (Val)</td>
<td>A (Ile)</td>
<td>1.00</td>
<td>2.11 (0.91–4.88)</td>
<td>1.19 (0.07–19.11)</td>
<td>0.117</td>
</tr>
<tr>
<td>CYP19A1-01</td>
<td>G (Val)</td>
<td>A (Val)</td>
<td>1.00</td>
<td>1.80 (1.06–3.04)</td>
<td>1.62 (0.95–2.74)</td>
<td>0.17</td>
</tr>
<tr>
<td>CYP19A1-14</td>
<td>T</td>
<td>C</td>
<td>1.00</td>
<td>1.98 (1.14–3.46)</td>
<td>1.79 (1.04–3.10)</td>
<td>0.109</td>
</tr>
<tr>
<td>GSTT1-02</td>
<td>present</td>
<td>null</td>
<td>1.00</td>
<td>1.71 (1.10–2.68)</td>
<td>1.08 (0.64–1.81)</td>
<td>0.688</td>
</tr>
<tr>
<td>HSD3B1-03</td>
<td>G (Ile)</td>
<td>A (Val)</td>
<td>1.00</td>
<td>2.50 (0.86–7.27)</td>
<td>1.47 (0.56–3.80)</td>
<td>0.576</td>
</tr>
<tr>
<td>IL10-02</td>
<td>C</td>
<td>A</td>
<td>1.00</td>
<td>1.43 (0.97–2.12)</td>
<td>1.24 (0.70–2.21)</td>
<td>0.181</td>
</tr>
<tr>
<td>IL10-03</td>
<td>G</td>
<td>A</td>
<td>1.00</td>
<td>1.68 (0.99–2.87)</td>
<td>1.51 (0.86–2.65)</td>
<td>0.283</td>
</tr>
<tr>
<td>IL1B-01</td>
<td>T</td>
<td>C</td>
<td>1.00</td>
<td>1.42 (0.87–2.30)</td>
<td>1.70 (1.01–2.85)</td>
<td>0.048</td>
</tr>
<tr>
<td>IL4-01</td>
<td>T</td>
<td>C</td>
<td>1.00</td>
<td>1.63 (0.96–2.74)</td>
<td>1.99 (1.21–3.28)</td>
<td>0.941</td>
</tr>
<tr>
<td>MPO-02</td>
<td>G</td>
<td>A</td>
<td>1.00</td>
<td>1.98 (0.81–4.82)</td>
<td>2.36 (0.99–5.63)</td>
<td>0.068</td>
</tr>
<tr>
<td>MTHFR-02</td>
<td>T (Val)</td>
<td>C (Ala)</td>
<td>1.00</td>
<td>2.30 (1.07–4.91)</td>
<td>2.21 (1.05–4.65)</td>
<td>0.175</td>
</tr>
<tr>
<td>MTRR-01</td>
<td>G</td>
<td>A</td>
<td>1.00</td>
<td>1.32 (0.83–2.10)</td>
<td>1.81 (1.09–3.02)</td>
<td>0.021</td>
</tr>
<tr>
<td>OGG1-04</td>
<td>C (Ser)</td>
<td>G (Cys)</td>
<td>1.00</td>
<td>0.68 (0.45–1.02)</td>
<td>0.41 (0.14–1.20)</td>
<td>0.019</td>
</tr>
<tr>
<td>PGK2-11</td>
<td>C</td>
<td>T</td>
<td>1.00</td>
<td>0.36 (0.11–1.11)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SOD2-01</td>
<td>C (Ala)</td>
<td>T (Val)</td>
<td>1.00</td>
<td>0.66 (0.42–1.05)</td>
<td>0.90 (0.53–1.50)</td>
<td>0.776</td>
</tr>
<tr>
<td>TGFBI-02</td>
<td>C (Thr)</td>
<td>T (Ile)</td>
<td>1.00</td>
<td>0.25 (0.08–0.74)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TP53-01</td>
<td>G (Arg)</td>
<td>C (Pro)</td>
<td>1.00</td>
<td>1.43 (0.96–2.15)</td>
<td>1.71 (1.00–2.95)</td>
<td>0.027</td>
</tr>
</tbody>
</table>

\(^1\)Determined by unconditional logistic regression.
## TABLE II

TOP SNP COMBINATIONS AND ASSOCIATED ORs FOR DETERMINING COLON CANCER RISK AS DETERMINED BY PIA WITH THE GINI AND % WRONG SCORING FUNCTIONS

<table>
<thead>
<tr>
<th>SNPs</th>
<th>Gini scoring function</th>
<th>% Wrong scoring function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top model</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td>1</td>
<td>TGFB1-02</td>
<td>n/a</td>
</tr>
<tr>
<td>2</td>
<td>IL1B-01, IL1B-03</td>
<td>1.42 (1.22–1.65)</td>
</tr>
<tr>
<td>3</td>
<td>IL1B-01, ALOX12-02, IL1B-03</td>
<td>1.47 (1.24–1.74)</td>
</tr>
<tr>
<td>4</td>
<td>WRN-03, IL5-02, IL10-02, CYP19A1-06</td>
<td>1.87 (1.56–2.23)</td>
</tr>
<tr>
<td>Rank</td>
<td>First order</td>
<td>Second order</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>--------------</td>
</tr>
<tr>
<td>1</td>
<td>TGFB1-02</td>
<td>ILIB-01</td>
</tr>
<tr>
<td>2</td>
<td>IL4-01</td>
<td>IL4-01</td>
</tr>
<tr>
<td>3</td>
<td>PTGS2-11</td>
<td>MTRR-01</td>
</tr>
<tr>
<td>4</td>
<td>GSTT1-02</td>
<td>MTRR-01</td>
</tr>
<tr>
<td>5</td>
<td>CYP19A1-14</td>
<td>TP53-01</td>
</tr>
<tr>
<td>6</td>
<td>CYP19A1-01</td>
<td>SOD2-01</td>
</tr>
<tr>
<td>7</td>
<td>OGG1-04</td>
<td>IL10-02</td>
</tr>
<tr>
<td>8</td>
<td>CASP8-03</td>
<td>CYP19A1-06</td>
</tr>
<tr>
<td>9</td>
<td>PTGS2-03</td>
<td>CYP19A1-09</td>
</tr>
<tr>
<td>10</td>
<td>GSTT1-02</td>
<td>MTRR-01</td>
</tr>
</tbody>
</table>

**TABLE III**

<table>
<thead>
<tr>
<th>Rank</th>
<th>First order</th>
<th>Second order</th>
<th>Third order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TGFB1-02</td>
<td>ILIB-01</td>
<td>ILIB-01</td>
</tr>
<tr>
<td>2</td>
<td>IL4-01</td>
<td>IL4-01</td>
<td>ILIB-01</td>
</tr>
<tr>
<td>3</td>
<td>PTGS2-11</td>
<td>MTRR-01</td>
<td>ILIB-01</td>
</tr>
<tr>
<td>4</td>
<td>GSTT1-02</td>
<td>MTRR-01</td>
<td>SOD2-01</td>
</tr>
<tr>
<td>5</td>
<td>CYP19A1-14</td>
<td>TP53-01</td>
<td>GSTT1-02</td>
</tr>
<tr>
<td>6</td>
<td>CYP19A1-01</td>
<td>SOD2-01</td>
<td>GSTT1-02</td>
</tr>
<tr>
<td>7</td>
<td>OGG1-04</td>
<td>IL10-02</td>
<td>ILIB-01</td>
</tr>
<tr>
<td>8</td>
<td>CASP8-03</td>
<td>CYP19A1-06</td>
<td>MTRR-01</td>
</tr>
<tr>
<td>9</td>
<td>PTGS2-03</td>
<td>CYP19A1-09</td>
<td>IL4-01</td>
</tr>
<tr>
<td>10</td>
<td>GSTT1-02</td>
<td>MTRR-01</td>
<td>SOD2-01</td>
</tr>
</tbody>
</table>

**Gini Cost Function**

<table>
<thead>
<tr>
<th>Score</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.489</td>
<td>0.493</td>
<td>0.446</td>
</tr>
<tr>
<td>0.490</td>
<td>0.491</td>
<td>0.471</td>
</tr>
<tr>
<td>0.490</td>
<td>0.491</td>
<td>0.471</td>
</tr>
<tr>
<td>0.490</td>
<td>0.491</td>
<td>0.471</td>
</tr>
<tr>
<td>0.490</td>
<td>0.491</td>
<td>0.471</td>
</tr>
<tr>
<td>0.490</td>
<td>0.491</td>
<td>0.471</td>
</tr>
<tr>
<td>0.490</td>
<td>0.491</td>
<td>0.471</td>
</tr>
<tr>
<td>0.490</td>
<td>0.491</td>
<td>0.471</td>
</tr>
<tr>
<td>0.490</td>
<td>0.491</td>
<td>0.471</td>
</tr>
<tr>
<td>0.490</td>
<td>0.491</td>
<td>0.471</td>
</tr>
<tr>
<td>Rank</td>
<td>Gini Cost Function</td>
<td>Score</td>
</tr>
<tr>
<td>------</td>
<td>-------------------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>WRN-03</td>
<td>0.39</td>
</tr>
<tr>
<td>2</td>
<td>IL4R-02</td>
<td>0.39</td>
</tr>
<tr>
<td>3</td>
<td>WRN-03</td>
<td>0.39</td>
</tr>
<tr>
<td>4</td>
<td>GSTP1-01</td>
<td>0.39</td>
</tr>
<tr>
<td>5</td>
<td>WRN-03</td>
<td>0.39</td>
</tr>
<tr>
<td>6</td>
<td>GSTP1-01</td>
<td>0.39</td>
</tr>
<tr>
<td>7</td>
<td>TGFB1-01</td>
<td>0.39</td>
</tr>
<tr>
<td>8</td>
<td>TP53-01</td>
<td>0.39</td>
</tr>
<tr>
<td>9</td>
<td>TGFB1-01</td>
<td>0.39</td>
</tr>
<tr>
<td>10</td>
<td>WRN-03</td>
<td>0.39</td>
</tr>
<tr>
<td>11</td>
<td>GSTP1-01</td>
<td>0.39</td>
</tr>
<tr>
<td>12</td>
<td>WRN-03</td>
<td>0.39</td>
</tr>
<tr>
<td>13</td>
<td>MTRR-01</td>
<td>0.39</td>
</tr>
<tr>
<td>14</td>
<td>GPX1-07</td>
<td>0.39</td>
</tr>
<tr>
<td>15</td>
<td>TP53-01</td>
<td>0.39</td>
</tr>
<tr>
<td>16</td>
<td>WRN-03</td>
<td>0.39</td>
</tr>
<tr>
<td>17</td>
<td>GSTP1-01</td>
<td>0.39</td>
</tr>
<tr>
<td>18</td>
<td>GPX1-06</td>
<td>0.39</td>
</tr>
<tr>
<td>19</td>
<td>WRN-03</td>
<td>0.39</td>
</tr>
<tr>
<td>20</td>
<td>GSTP1-01</td>
<td>0.39</td>
</tr>
<tr>
<td>SNP-SNP model</td>
<td>1st order</td>
<td>2nd order</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>IL4-01</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MTRR-01</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>TP53-01</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>GSTT1-02</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>SOD2-01</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>WRN-03</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>IL1B-01</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>IL10-02</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>GSTP1-01</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>CYP19A0-06</td>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>CYP19A1-14</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>CYP19A8-09</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>ESR1-03</td>
<td>34</td>
<td>36</td>
</tr>
<tr>
<td>CYP19A0-01</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>IL6-03</td>
<td>19</td>
<td>24</td>
</tr>
<tr>
<td>IL8-01</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>IL4-03</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>HSD17B1-03</td>
<td>30</td>
<td>26</td>
</tr>
<tr>
<td>GSTD1-03</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td>CYP17A0-05</td>
<td>36</td>
<td>35</td>
</tr>
<tr>
<td>GPX1-01</td>
<td>32</td>
<td>31</td>
</tr>
<tr>
<td>MTHFR-01</td>
<td>25</td>
<td>17</td>
</tr>
<tr>
<td>CYP19A0-04</td>
<td>35</td>
<td>26</td>
</tr>
<tr>
<td>IL4-03</td>
<td>30</td>
<td>22</td>
</tr>
<tr>
<td>HSD17B1-03</td>
<td>30</td>
<td>24</td>
</tr>
<tr>
<td>ALOX12-02</td>
<td>48</td>
<td>4</td>
</tr>
<tr>
<td>ALOX5-01</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>TGFBI-01</td>
<td>46</td>
<td>33</td>
</tr>
<tr>
<td>EGF-08</td>
<td>14</td>
<td>28</td>
</tr>
<tr>
<td>XRC-01</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>CYP17A0-01</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>IL13-03</td>
<td>44</td>
<td>33</td>
</tr>
<tr>
<td>OGG1-04</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>MTHFR-02</td>
<td>10</td>
<td>41</td>
</tr>
<tr>
<td>PTGS2-11</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>NQO1-01</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>GPX1-07</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>CASP8-03</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>SHBG-01</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>NBS1-01</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>MPO-02</td>
<td>31</td>
<td>7</td>
</tr>
<tr>
<td>HSD17B3-01</td>
<td>27</td>
<td>16</td>
</tr>
<tr>
<td>ERCC2-02</td>
<td>44</td>
<td>37</td>
</tr>
<tr>
<td>RACE</td>
<td>21</td>
<td>28</td>
</tr>
<tr>
<td>GPX1-06</td>
<td>43</td>
<td>37</td>
</tr>
<tr>
<td>EPHX1-01</td>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>TNF-02</td>
<td>7</td>
<td>33</td>
</tr>
<tr>
<td>TGFBI-02</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>IL6-01</td>
<td>46</td>
<td>41</td>
</tr>
<tr>
<td>SNP-SNP model</td>
<td>1st order</td>
<td>2nd order</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>BRCA2-01</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>SRD5A2-01</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>ESR1-07</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>ALOX5-07</td>
<td>37</td>
<td>43</td>
</tr>
<tr>
<td>ERCC5-02</td>
<td>39</td>
<td>47</td>
</tr>
<tr>
<td>GPX1-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBC3-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VEGF-04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLAB-02</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>CHEKI-02</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>NOS2A-02</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>MAOA-01</td>
<td>33</td>
<td>38</td>
</tr>
<tr>
<td>CAT-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRD2-01</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>RAD23B-04</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>EDN1-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC25C-01</td>
<td>32</td>
<td>47</td>
</tr>
<tr>
<td>WRN-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOS3-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP19A1-08</td>
<td>40</td>
<td>46</td>
</tr>
<tr>
<td>ESR1-08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDKN1A-04</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>IL4R-03</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>CDKN2A-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC25A-02</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>