IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 27, 2019, accepted September 16, 2019, date of publication September 25, 2019, date of current version October 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2943614

Detecting SNP–SNP Interactions in Imbalanced Case-Control Study

CHENG-HONG YANG[1,2], (Senior Member, IEEE), LI-YEH CHUANG[3], AND YU-DA LIN[1], (Member, IEEE)

[1]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung 807, Taiwan
[2]Ph.D. Program in Biomedical Engineering, Kaohsiung Medical University, Kaohsiung 807, Taiwani
[3]Department of Chemical Engineering, Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung 807, Taiwan

Corresponding authors: Li-Yeh Chuang (chuang@isu.edu.tw) and Yu-Da Lin (yudalinemail@gmail.com)

This work was supported in part by the Ministry of Science and Technology, China, under Grant 108-2221-E-992-031-MY3, Grant 108-2221-E-214-019-MY3, and Grant 108-2811-E-992-502.

ABSTRACT SNP–SNP interactions are particularly informative biomarkers regarding the genetic components of disease risk. However, SNP–SNP interaction identifications are yet limited in imbalanced case–control study. In this study, we proposed a multiobjective multifactor dimensionality reduction (MOMDR) based on three balancing approaches (BMOMDR), including (1) stratified K-fold cross-validation; (2) balanced estimation of ratio between cases and controls; (3) balanced measures of SNP–SNP interactions, to effectively identify SNP–SNP interaction in imbalanced case–control study. BMOMDR was evaluated by extensive experiments on both simulated imbalanced case–control datasets and real genome-wide data from Wellcome Trust Case Control Consortium (WTCCC). For the simulated datasets, the results indicated that three balancing approaches can enhance the detection success rate of SNP–SNP interaction by MOMDR in imbalanced datasets. For WTCCC datasets, the results of SNP–SNP interaction detection obtained from BMOMDR revealed statistically significant (p <0.0001), revealing that BMOMDR can effectively identify SNP–SNP interaction in imbalanced case–control study. BMOMDR is freely available at http://shorturl.at/bluJS.

INDEX TERMS SNP–SNP interactions, multiobjective approach multifactor dimensionality reduction, imbalanced case-control study.

I. INTRODUCTION

Genome-wide association studies (GWASs) have demonstrated that multilocus single-nucleotide polymorphisms (SNPs) influence some diseases [1]–[5]. SNP–SNP interactions might be involved in some complex traits of these diseases [6]–[8], and the determination of SNP–SNP interactions could resolve missing heritability concerns [9]. To improve genetic association studies, the development of efficient analysis methods for SNP–SNP interactions is a key concern [10], [11].

Model-free approaches have been employed in the detection of SNP–SNP interactions; such approaches are not obligated to hypothesize about genetic models and data [12]–[14]. A well-known model-free approach, multifactor dimensionality reduction (MDR), was applied in a case–control study [15]. MDR is able to consider the entire set of predictive rules for any multilocus (m-locus) combination. It can reduce the dimensionality of any m-locus combination by taking a high-dimensional m-locus space as input and returning a one-dimensional space as output. As has been previously reported in the literature, SNP–SNP interaction was evaluated using correct classification rate (CCR) with two-way contingency tables and cross validation (CV), so that overfitting of the training data can be avoided and false positive errors can be minimized [15]. MDR has exhibited satisfactory performance with nonlinear effects and high dimension datasets; multiple case–control studies have successfully conducted by MDR to explain the conditions, including oral cancer [16], hypertension [17], and breast cancer [18].

Although MDR has exhibited numerous advantages in SNP–SNP interaction identification, original MDR does have its limitations, such as poor certainty in the multifactor category, calculation costs, and specific data issues [19].

The associate editor coordinating the review of this manuscript and approving it for publication was Corrado Mencar.

Based on the limitations of MDR, it can be improved in three aspects. The first area focuses on the effective classification of multiple factors into risk groups through using various technology such as odds ratio-based MDR [20], log-linear model-based MDR [20], and multiobjective (MO)-based MDR (MOMDR) [21]. The term ''MO method'' denotes a multiple-criteria decision analysis technique; a decision problem with multiple conflicting criteria can be solved by an MO method [22]. An MO approach is able to evaluate a variety of measures at the same time and return the appropriate solution as output [23]. The second aspect is using the technology to reduce the calculation cost of analysis, including the unified model-based MDR [24], fast MDR [25], graphical processing unit–based MDR [26], and differential evolution (DE)-based MDR [27]. The third aspect is applying MDR to quantitative traits, survival data, and imbalanced datasets by using techniques such as quantitative MDR [28], Cox-MDR [29], MDR-based balanced CCR (bCCR) [30], and MDR-based adjusting the ratio in risk classes and classification errors [31]. However, most of the techniques applied to imbalanced case–control study still involve using the standard MDR with the CCR measure [30], [31]. Alternative contingency table measures [32] and MOMDR [21] have not been addressed in imbalanced case–control study yet.

Due to MDR incorrectly detects SNP–SNP interactions in imbalanced datasets [30], resampling approaches are typically applied to overcome the limitations of MDR in imbalanced case–control study. However, some of the crucial information may be lost because some samples may be undersampled. Yang *et al*. proposed an MO technique to enable MDR to simultaneously take various approaches for the identification of potential SNP–SNP interactions [21]. However, the performance of MOMDR [21] on imbalanced case–control datasets has not been addressed. The aim of this study was to improve the performance of MDR and MOMDR in imbalanced case–control study. Here, we proposed a MOMDR-based balancing approaches (BMOMDR) which adopted three balancing approaches, including (1) stratified K-fold cross-validation; (2) balanced estimation of ratio between cases and controls; (3) balanced measures of SNP–SNP interactions, to overcome MDR and MOMDR limitations in imbalanced case–control study. We experimented with various imbalance case–control data and used a GWAS data to evaluate the performance of BMOMDR.

II. METHODS

A. DEFINITION OF SNP-SNP INTERACTION DETECTION

For SNP–SNP interaction detection, an SNP S consists of a set of values $\{1, 2, 3\}$ which are three genotypes, including the homozygous reference genotype, the heterozygous genotype, and the homozygous variant genotype. s represents an index–value (i, v) pair, where i is an index and v is some element of $\{1, 2, 3\}$. A predictive rule relates features and class variables (i.e., cases and controls). A predictive rule formalizes an epistatic phenomenon; for some conjunction of n literals $(s_1, s_2, \ldots, s_n)$ written as r and for some class label ζ ($\zeta = 0$ denotes controls, $\zeta = 1$ denotes cases) there exists some literals of conjunctive rules (r, ζ): $s_1 \cap s_2 \cap \ldots \cap s_n \rightarrow \zeta$. Therefore, a SNP–SNP interaction of order m is represented as an m-locus combination having 3^m predictive rules in the form $X = (r_1, r_2, \ldots, r_n)$, where $n = 3^m$, $\forall r \in S_1 \cup S_2 \cup \ldots \cup S_m$, and where ''3'' represents the genotype type. To detect SNP–SNP interactions, one must identify the m-locus combination that has the maximum quality according to some statistical measure.

B. MOMDR

In MOMDR, suppose the m-locus combination X is a decision vector. Next, suppose some set of measures $f_1, f_2, \ldots, f_i$ are objective functions. In an MO maximization, an objective function can be constructed to obtain a set of values from multiple measures as

$$\text{maximize} \begin{cases} f_1(x) = measure_1(x) \\ f_2(x) = measure_2(x), \end{cases} \tag{1}$$

where functions f_1 and f_2 are the objective functions for the MDR measures. A set of j feasible m-locus combinations is defined by some vector $X^* = (X_1, X_2, \ldots, X_j)$. X_1 dominates another solution X_2 if $f_p(X_1) \geq f_p(X_2)$ for $p = 1, 2, \ldots, i$. A solution vector X is deemed Pareto optimal when no other solution dominates X. The Pareto set X^* consists of every solution vector $X \in X^*$ that is Pareto optimal.

A Pareto set operation is introduced to extend MDR to use multiple measures simultaneously to assess the quality of an m-locus combination. Extra storage space is produced by the Pareto set operation. Pareto set filter operators choose candidates from m-locus combinations. A total of K Pareto sets are obtained through the K-fold cross-validation (CV). In MOMDR, the Pareto cross-validation consistency (CVC) operation is used to determine the multiple solutions depending on the most instances of solutions in K Pareto sets. MOMDR process comprises these stages: (1) The dataset is divided into K subsets for the CV calculation and generate K Pareto sets. (2) All feasible m-locus combinations are generated, and data reduction technique is used to divide multilocus genotypes as high- and low-risk groups. The set of 3^m predictive rules is transformed into a two-way contingency table. (3) The m-locus combinations are evaluated using multiple measures. (4) The K Pareto set are respectively updated according to the multiple measures in K-fold CV. (5) The m-locus combinations from all Pareto sets are counted. The m-locus combinations with highest CVC are regarded as the optimal solutions.

C. BMOMDR

BMOMDR adopted three balancing approaches, including (1) stratified K-fold CV datasets; (2) balanced estimation of ratio between cases and controls; (3) balanced measures. The approaches can improve the ability of MOMDR to detect SNP–SNP interactions. BMOMDR consists of the following titles:

Algorithm 1 Stratified Random K-Fold CV

01: Divide the samples into cases and controls.
02: Randomly shuffle the samples in each set.
03: Count the total number of samples in the case set (cases) and the total number of samples in the control set (controls).
04: Compute the ratio between cases and controls.
05: Classify the samples of cases and controls into a j^{th}-fold CV subset according to the ratio between cases and controls, in which j is the index of the CV subset.

Step 1: The dataset is assigned into K subsets of CV. The CV uses the stratified random K-fold [33], and the pseudo-code is shown in Algorithm 1. The stratified random K-fold enables each fold to have the same proportion of cases and controls.

Step 2: Feasible m-locus combinations are calculated.

Step 3: The feasible m-locus combination constructs a table containing 3^m predictive rules. The samples of training datasets are classified into the corresponding predictive rules, and the cases and controls in the predictive rules are counted.

Step 4: Evaluate all case–control ratios for predictive rules by balanced estimation of ratio between cases and controls (2).

$$\begin{aligned} f_{ratio}(X) &= \{\theta_a\} \\ s.t.\theta_a &= \frac{n_{+0} \times n_{a1}}{n_{+1} \times n_{a0}} \end{aligned} \tag{2}$$

where X is the set of 3^m predictive rules in an m-locus combination. a is the index of predictive rule. n_{a0} and n_{a1} are the numbers of samples in the a^{th} predictive rule in the control group and case group, respectively. n_{+0} and n_{+1} are the total number of samples in the control group and case group, respectively. Subsequently, the high- and low-risk groups of predictive rules are determined in X. The a^{th} predictive rule is labeled as belonging to the high-risk group when θ_a is larger than 1 [15]; otherwise, it belongs to the low-risk group.

Step 5: A two-way contingency table is calculated from 3^m-labeled predictive rules on the basis of their grouping and outcome. The *TP* (true positive), *FP* (false positive), *FN* (false negative), and *TN* (true negative) of the two-way contingency table are calculated according to the number of samples belonging to the corresponding groups and outcomes, and these values are calculated using (3).

$$\begin{cases} TP = \sum\limits_{a \in \{\theta_a, \theta_a \geq 1\}} t_{a1} \\ FP = \sum\limits_{a \in \{\theta_a, \theta_a \geq 1\}} t_{a0} \\ FN = \sum\limits_{a \in \{\theta_a, \theta_a < 1\}} t_{a1} \\ TN = \sum\limits_{a \in \{\theta_a, \theta_a \geq 1\}} t_{a0} \end{cases} \tag{3}$$

where t_{ab} is the set of individual matches to the a^{th} multifactor class in the b outcome status, where $b = 1$ for the case group and $b = 0$ for the control group.

Step 6: The objective function of an m-locus combination X is calculated. Here, a balancing technique was used to improve the *TP*, *FP*, *FN*, and *TN* for handling unbalanced datasets to calculate the proportion of appropriately classified individuals. *bTP* (balanced true positive rate), *bFP* (balanced false positive rate), *bFN* (balanced false negative rate), and *bTN* (balanced true negative rate) are formulated as the balanced *TP*, *FP*, *FN*, and *TN* values, respectively (4).

$$\begin{cases} bTP = \dfrac{TP}{TP + FN} \\ bFP = \dfrac{FP}{FP + TN} \\ bFN = \dfrac{FN}{TP + FN} \\ bTN = \dfrac{TN}{FP + TN} \end{cases} \tag{4}$$

Objective 1: We used *bCCR*, which was proposed by Velez *et al*. [30] and calculated using *bTP* and *bTN*:

$$\begin{aligned} f(X) &= \text{measure}(X) \\ &= bCCR(X) \\ &= 0.5 \times (bTP + bTN) \end{aligned} \tag{5}$$

$$\begin{aligned} f(X) &= measure(X) \\ &= bNMI(X) = \frac{H(y) - H(y|x)}{H(y)} = \frac{2\left\{\begin{array}{l} bPN\,log_2 bPN + bTP log_2 bTP + bFN log_2 bFN \\ +bTN log_2 bTN + bFP log_2 bFP - bP log_2 bP \\ -bTPN log_2 bTPN - bTNP log_2 bTNP - bN log_2 bN \end{array}\right\}}{2\left\{bPN\,log_2 bPN - bTPN log_2 bTPN - bTNP log_2 bTNP\right\}} \end{aligned} \tag{6}$$

$$\text{where} \begin{cases} bPN = bTP + bFP + bFN + bTN \\ bP = bTP + bFP \\ bN = bTN + bFN \\ bTPN = bTP + bFN \\ bTNP = bTN + bFP \end{cases}$$

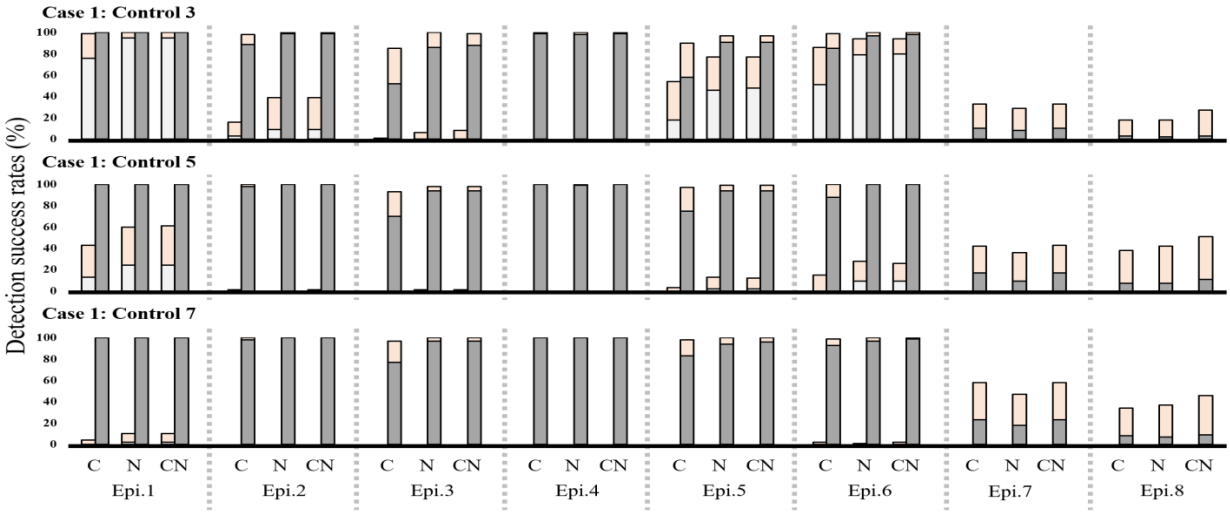

FIGURE 1. Comparison of detection success rates between MDR, MOMDR, BMDR, and BMOMDR in eight models with marginal effects in imbalanced case–control study. The C bar represents the results of MDR (CCR, left bar) and BMDR (balanced CCR, right bar). The N bar provides the results of MDR (NMI, left bar) and BMDR (balanced NMI, right bar). The CN bar presents the results of MOMDR (CCR and NMI, left bar) and BMOMDR (balanced CCR and NMI, right bar). In each bar, the lower region (gray and dark gray) represents the detection success rate for CVC/ Pareto CVC = 5; the upper region (yellow) represents the detection success rate for CVC/Pareto CVC < 5. The absence of bars indicates zero detection success rate. The dataset comprised 1,000 SNPs, and the sample sizes were 1,600 (400 cases and 1,200 controls, 1:3), 2,400 (400 cases and 2,000 controls, 1:5), and 3,200 (400 cases and 2,800 controls, 1:7). Under each setting, the detection success rate was calculated as the proportion for 100 datasets in which the specific disease-associated SNP–SNP interactions were detected.

The *bCCR* value is in the 0–1 interval, with 1 expressing the optimal solution.

Objective 2: The normalized mutual information (*NMI*) measure has been applied to MDR [32]. We modified the *NMI* measure by using a balancing technique, and the modified measure (*bNMI*) can be expressed as (6), as shown at the bottom of the previous page. The maximum value expresses the optimal solution.

Step 7: Pareto operation. In each CV, the all non-dominated candidates are added into the Pareto set.

Step 8: Pareto CVC calculation. In each fold CV, steps 3–7 are repeated until all m-locus combinations have been evaluated. Thus, the K Pareto sets can be obtained. In the Pareto CVC operation, the candidate with the highest number of occurrences in the K Pareto sets is regarded as the optimal result; if numerous candidates occur equally as frequently, these candidates are regarded as optimal results. For CVC $= K$, we interpret that the SNP-SNP interactions considered were those that appeared K times among the K-fold CV.

III. RESULTS

We defined BMOMDR as MOMDR using both *bCCR* and *bNMI* estimations; MOMDR as MOMDR using both *CCR* and *NMI* estimations; BMDR-CCR and BMDR-NMI as MDR using the *bCCR* and *bNMI* estimations respectively; and MDR-CCR and MDR-NMI as MDR using *CCR* and *NMI* estimations respectively. BMDR-CCR and BMOMDR used the imbalanced CCR version proposed by Velez *et al*. [30]. The performance of BMOMDR was evaluated by comparing it with that of MOMDR, BMDR-CCR, BMDR-NMI, MDR-CCR, and MDR-NMI by using simulated imbalanced case–control datasets. A GWAS dataset obtained from the Wellcome Trust Case Control Consortium (WTCCC) was used to evaluate the performance of BMOMDR [34].

A. SNP–SNP INTERACTION DETECTION FOR IMBALANCED CASE–CONTROL DATASETS

1) TWO-LOCUS DISEASE MODEL WITH MARGINAL EFFECTS IN IMBALANCED CASE–CONTROL STUDY

The performance levels of MDR-CCR, MDR-NMI, BMDR-CCR, BMDR-NMI, MOMDR, and BMOMDR were characterized in terms of disease loci with marginal effects in imbalanced case–control study. Eight disease models with marginal effects were taken [35], [36]. Supplementary Table S1 shows the multilocus penetrances of eight disease models. GAMETES software was used to generate imbalanced case–control datasets [37]. Each disease model generated 100 datasets. In each disease model, we analyzed three imbalanced category in which the ratios of cases to controls were 1:3, 1:5, and 1:7. Each dataset had its own correct solution as an interacting SNP pair, and other SNPs were generated at minor allele frequencies (MAFs) selected uniformly from [0.05, 0.5). The success rates of detection expressed the frequencies of correct answer detection within the 100 datasets.

The detection success rates of MDR-CCR, MDR-NMI, BMDR-CCR, BMDR-NMI, MOMDR, and BMOMDR are presented in Fig. 1. The detection success rates of both BMDR-CCR and BMDR-NMI were higher than those of MDR in the datasets with case–control ratios of 1:3, 1:5, and 1:7. For the detection of SNP–SNP interactions with

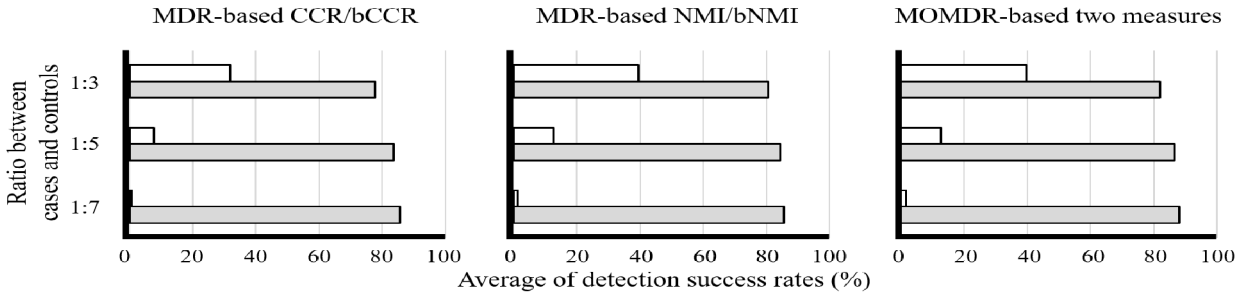

FIGURE 2. Comparison of the average detection success rates in eight models with marginal effects in imbalanced case–control study. White bars indicate the average detection success rates of MDR-based and MOMDR-based original measures, and gray bars present the average detection success rates of MDR-based and MOMDR-based balanced measures. The 1:3 ratio indicates a total of 1,600 samples (400 cases, 1,200 controls); the 1:5 ratio indicates a total of 2400 samples (400 cases, 2,000 controls); the 1:7 ratio indicates a total of 3,200 samples (400 cases, 2,800 controls).

TABLE 1. BMDR or BMOMDR compared with MDR or MOMDR for detection success rate in ONLY CVC = 5 of two-locus disease model with marginal effects in imbalanced case–control study using Wilcoxon signed-rank test.

	R^-	R^+	$R^=$	*P* value
400 cases and 1,200 controls, 1:3				
BMDR-CCR vs MDR-CCR	0	8	0	**0.012**
BMDR-NMI vs MDR-NMI	0	8	0	**0.012**
BMOMDR vs MOMDR	0	8	0	**0.012**
400 cases and 2,000 controls, 1:5				
BMDR-CCR vs MDR-CCR	0	8	0	**0.012**
BMDR-NMI vs MDR-NMI	0	8	0	**0.012**
BMOMDR vs MOMDR	0	8	0	**0.012**
400 cases and 2,800 controls, 1:7				
BMDR-CCR vs MDR-CCR	0	8	0	**0.012**
BMDR-NMI vs MDR-NMI	0	8	0	**0.012**
BMOMDR vs MOMDR	0	8	0	**0.012**

R^-: negative ranks, R^+: positive ranks, $R^=$: ties, bold type indicates the significant improvement ($P < 0.05$).

CVC = 5, both BMDR-CCR and BMDR-NMI achieved superior detection success rates to those of MDR in eight disease models. Moreover, BMOMDR achieved detection success rates superior to those of MOMDR in all tests, and with Pareto CVC = 5 BMOMDR outperformed MOMDR. The performance of BMOMDR was evaluated using a Wilcoxon signed-rank test, in which p <0.05 indicated significant superiority of BMDR (BMDR-CCR and BMDR-NMI) or BMOMDR relative to MDR or MOMDR. As presented in Table 1, BMDR-CCR, BDR-NMI, and BMOMDR exhibited significantly superior detection success rates to those of MDR and MOMDR, respectively (p <0.05).

The average detection success rates in eight models with marginal effects are presented in Fig. 2. The average detection success rates for 1:3, 1:5, and 1:7 datasets were 32.00, 7.75, and 0.75, respectively, in MDR-CCR; 39.50, 12.75, and 1.38, respectively, in MDR-NMI; and 39.75, 12.63, and 1.50, respectively, in MOMDR. The detection success rates of MDR-CCR, MDR-NMI, and MOMDR decreased as the case–control ratio increased. The average detection success rates for 1:3, 1:5, and 1:7 were 77.88, 83.75, and 85.75, respectively, in BMDR-CCR; 80.50, 84.38, and 85.50, respectively, in BMDR-NMI; and 82.00, 86.38, and 88.00, respectively, in BMOMDR. The detection success rates of BMDR-CCR, BMDR-NMI, and BMOMDR increased with the case–control ratio, revealing that BMDR-CCR, BMDR-NMI, and BMOMDR can effectively detect SNP–SNP interactions in disease loci with marginal effects in imbalanced case–control study.

2) TWO-LOCUS DISEASE MODEL WITHOUT MARGINAL EFFECTS IN IMBALANCED CASE–CONTROL STUDY

MDR-CCR, MDR-NMI, BMDR-CCR, BMDR-NMI, MOMDR, and BMOMDR were assessed with 60 two-locus and pure models without marginal effects in imbalanced case–control study [38]. Details on the 60 disease models of multilocus penetrances are provided in Supplementary Tables S2-S7. The phenotypic variations of all diseases were controlled by h^2 values that were greater than or equal to 0.025 and less than or equal to 0.2, with MAFs of 0.2 as well as 0.4. For each model, 100 datasets containing 1000 SNPs were randomly generated using GAMETES, of which two SNPs were specific SNP pair, and other SNPs were uniformly chosen from [0.05, 0.5) MAFs. For this analysis, three imbalanced datasets were selected in which the case–control ratios were 1:3, 1:5, and 1:7. The success rate of detection was determined by noting the frequency of a particular SNP pair detected in 100 datasets.

The 60 disease models without marginal effects were applied to characterize the performance of MDR-CCR, MDR-NMI, BMDR-CCR, BMDR-NMI, MOMDR, and BMOMDR in detecting SNP pairs. BMDR-CCR, BMDR-NMI, and BMOMDR achieved an improvement in performance compared with MDR and MOMDR in the disease models without marginal effects in imbalanced case–control study (Supplementary Fig. S1). In 45 of the 60 disease models, BMDR-CCR and BMDR-NMI had higher detection success rates than did MDR for datasets with 1:3, 1:5, and 1:7 ratios. In the remaining 15 models (data not shown), all algorithms obtained 100% detection success rates. In 45 disease models, BMOMDR had superior detection success rates compared with MOMDR. BMOMDR achieved higher

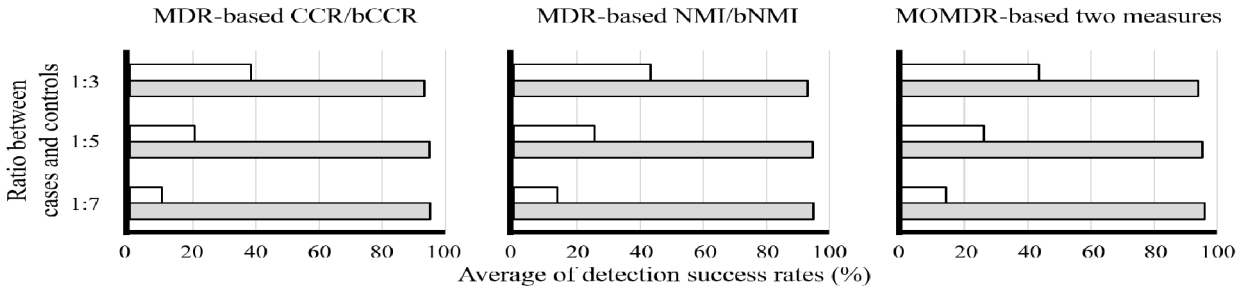

FIGURE 3. Comparison of the average detection success rates in 60 models without marginal effects in imbalanced case–control study. The white bar indicates the average detection success rates of MDR-based and MOMDR-based original measures, and the gray bar denotes the average detection success rates of MDR-based and MOMDR-based balanced measures. For the 1:3 ratio, the sample size was 1,600 (400 cases and 1,200 controls); for the 1:5 ratio, the sample size was 2,400 (400 cases and 2,000 controls); for the 1:7 ratio, the sample size was 3,200 (400 cases and 2,800 controls).

detection success rates with Pareto CVC = 5 than MOMDR. The performance of MDR (MDR-CCR and MDR-NMI), BMDR (BMDR-CCR and BMDR-NMI), MOMDR, and BMOMDR was evaluated using Wilcoxon signed-rank test in the 45 disease models without marginal effects. A p-value of <0.05 indicated significant superiority of BMDR (BMDR-CCR and BMDR-NMI) or BMOMDR compared with MDR (MDR-CCR and MDR-NMI) or MOMDR. For datasets of disease models without marginal effects in imbalanced case–control study, BMDR (BMDR-CCR and BMDR-NMI) and BMOMDR exhibited significantly higher detection success rates than did MDR (MDR-CCR and MDR-NMI) and MOMDR (p <0.05; Table 2).

TABLE 2. BMDR or BMOMDR compared with MDR or MOMDR for detection success rate in ONLY CVC = 5 of two-locus disease model without marginal effects in imbalanced case–control study using Wilcoxon signed-rank test.

	R^-	R^+	$R^=$	*P* value
400 cases and 1,200 controls, 1:3				
BMDR-CCR vs MDR-CCR	2	42	1	**<0.001**
BMDR-NMI vs MDR-NMI	1	37	7	**<0.001**
BMOMDR vs MOMDR	0	41	4	**<0.001**
400 cases and 2,000 controls, 1:5				
BMDR-CCR vs MDR-CCR	0	45	0	**<0.001**
BMDR-NMI vs MDR-NMI	0	45	0	**<0.001**
BMOMDR vs MOMDR	0	45	0	**<0.001**
400 cases and 2,800 controls, 1:7				
BMDR-CCR vs MDR-CCR	0	45	0	**<0.001**
BMDR-NMI vs MDR-NMI	0	44	1	**<0.001**
BMOMDR vs MOMDR	0	45	0	**<0.001**

R^-: negative ranks, R^+: positive ranks, $R^=$: ties, bold type indicates the significant improvement ($P < 0.05$).

The average detection success rates in the 60 models without marginal effects in imbalanced case–control study are illustrated in Fig. 3. The average detection success rates for 1:3, 1:5, and 1:7 were 38.49, 20.71, and 10.24, respectively, in MDR-CCR; 43.29, 25.51, and 13.87, respectively, in MDR-NMI; and 43.69, 26.11, and 14.31, respectively, in MOMDR. The detection success rates of both MDR (MDR-CCR and MDR-NMI) and MOMDR decreased as the case–control ratio increased. The average detection success rates for 1:3, 1:5, and 1:7 were 93.31, 94.96, and 95.22, respectively, in BMDR-CCR; 93.02, 94.60, and 94.82, respectively, in BMDR-NMI; and 93.91, 95.33, and 95.93, respectively, in BMOMDR. The detection success rates of both BMDR (BMDR-CCR and BMDR-NMI) and BMOMDR increased with the case–control ratio. When the datasets were imbalanced, the results indicated that BMDR (BMDR-CCR and BMDR-NMI) and BMOMDR could effectively detect SNP–SNP interactions at the disease loci without marginal effects in imbalanced case–control study.

B. SNP–SNP INTERACTION DETECTION FOR WTCCC DATASET

The performance of BMOMDR for large datasets was tested with a WTCCC dataset of patients who claimed to be white Europeans. WTCCC had collected information regarding 1,988 patients with coronary artery disease (CAD) from the United Kingdom and 1,500 controls (imbalanced case–control study). The individuals were genotyped with an Affymetrix GeneChip 500K Mapping Array Set. Overall, 500,569 SNPs were detected in GWAS [34]. BMOMDR was implemented the SNP–SNP interaction detection over all two-way combinations based on 5-fold CV in the n SNPs with 3,488 samples, where n is the number of SNPs in GWAS.

The results of SNP–SNP interaction detection using BMOMDR are listed in Table 3. The SNP–SNP interaction related genes were obtained from the dbSNP database of National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/snp/). Chromosomes include multiple detected SNP–SNP interactions; this is because MO allows BMOMDR to obtain multiple solutions. The raw datasets were subjected to chi-squared testing; all p values indicated the significance level of a SNP–SNP interaction between some pair of SNPs. The detected SNP–SNP interactions were all highly significant (p <0.0001) for a SNP–SNP interaction between the two relevant SNPs. The false positive rate (FPR) and false negative rate (FNR) were shown in Table 3. The FPR refers to the expectancy of the false positive ratio. FNR assumes that there are conditions to be checked, the conditional probability of negative test results is obtained. For each SNP–SNP interaction, the frequency of

TABLE 3. Summary of BMOMDR results for CAD based on WTCCC data.

Location[a]	SNP Groups	Related Genes	bCCR	bNMI	FPR[b]	FNR[c]	Pareto CVC	*p*-value	Times[d] (h)
Chr1	rs41399650, rs2999538	UNKNOWN, LOC105371442	0.788	0.356	0.418	0.241	5	<0.0001	16.6
	rs41399650, rs17163057	UNKNOWN, UNKNOWN	0.798	0.312	0.342	0.216	5	<0.0001	
Chr2	rs41453947, rs41509345	UNKNOWN, NCKAP5	0.798	0.281	0.260	0.186	5	<0.0001	18.2
Chr3	rs10866051, rs9874734	LOC105376942, LOC105374258	0.839	0.395	0.339	0.204	5	<0.0001	13.4
	rs10866051, rs17042882	LOC105376942, PLCL2	0.832	0.441	0.331	0.200	5	<0.0001	
Chr4	rs41426946, rs41529544	PPA2, UNKNOWN	0.810	0.309	0.263	0.183	5	<0.0001	12.6
Chr5	rs2201044, rs7443778	LOC105379160, LOC105379160	0.693	0.138	0.106	0.139	5	<0.0001	12.3
	rs41493746, rs41421845	UNKNOWN, LOC105374731	0.704	0.127	0.237	0.217	5	<0.0001	
Chr6	rs12198616, rs41489047	BTBD9, ADGRB3	0.788	0.286	0.346	0.221	5	<0.0001	12.3
	rs3006172, rs41489047	WDR27, ADGRB3	0.792	0.279	0.323	0.221	5	<0.0001	
Chr7	rs41437948, rs7777155	POU6F2, ZNF92	0.684	0.108	0.241	0.230	5	<0.0001	5.5
Chr8	rs41343444, rs41346747	CPQ, ST18	0.738	0.199	0.000	0.000	5	<0.0001	6.7
	rs35120859, rs17480050	UNKNOWN, CSGALNACT1	0.754	0.196	0.223	0.188	5	<0.0001	
Chr9	rs41354745, rs41424148	KANK1, UNKNOWN	0.726	0.192	0.048	0.056	5	<0.0001	4.7
	rs41424148, rs41502748	UNKNOWN, LOC107987122	0.719	0.193	0.564	0.328	5	<0.0001	
Chr10	rs41370151, rs2944490	FAM107B, TCERG1L	0.791	0.329	0.390	0.232	5	<0.0001	8.4
Chr11	rs41535846, rs41518446	UNKNOWN, MAML2	0.676	0.092	0.344	0.274	5	<0.0001	5.5
Chr12	rs16926425, rs7299571	SOX5, UNKNOWN	0.959	0.770	0.011	0.009	5	<0.0001	5.2
Chr13	rs9540728, rs7328649	PCDH9, FAM155A	0.825	0.344	0.101	0.093	5	<0.0001	3.6
Chr14	rs41324950, rs41491051	LOC105370603, SLC35F4	0.804	0.307	0.092	0.090	5	<0.0001	2.9
	rs41324950, rs41453247	LOC105370603, UNKNOWN	0.808	0.302	0.130	0.116	5	<0.0001	
Chr15	rs41461246, rs41418548	UBE2Q2P1, SHC4	0.702	0.123	0.346	0.258	5	<0.0001	2.3
Chr16	rs235633, rs41483646	UNKNOWN, UNKNOWN	0.768	0.224	0.304	0.215	5	<0.0001	2.0
Chr17	rs3785579, rs9892020	CACNG1, UNKNOWN	0.879	0.527	0.228	0.149	5	<0.0001	1.5
	rs3785579, rs7219778	CACNG1, SDK2	0.878	0.534	0.237	0.152	5	<0.0001	
	rs3785579, rs4795043	CACNG1, UNKNOWN	0.878	0.540	0.239	0.154	5	<0.0001	
	rs3785579, rs180171	CACNG1, LOC107983956	0.877	0.545	0.242	0.154	5	<0.0001	
	rs3785579, rs1870998	CACNG1, UNKNOWN	0.877	0.543	0.243	0.155	5	<0.0001	
	rs3785579, rs7226189	CACNG1, UNKNOWN	0.876	0.547	0.246	0.157	5	<0.0001	
	rs3785579, rs9912522	CACNG1, UNKNOWN	0.878	0.532	0.235	0.152	5	<0.0001	
Chr18	rs41470446, rs3794931	UNKNOWN, ZNF516	0.756	0.202	0.197	0.173	5	<0.0001	2.1
	rs4799934, rs3794931	CELF4, ZNF516	0.746	0.292	0.502	0.276	5	<0.0001	
Chr19	rs11671119, rs375299	BORCS8-MEF2B, UNKNOWN	0.579	0.079	0.803	0.378	5	<0.0001	0.9
	rs375299, rs41370444	UNKNOWN, UNKNOWN	0.641	0.062	0.477	0.322	5	<0.0001	
Chr20	rs2748666, rs41405046	UNKNOWN, UNKNOWN	0.884	0.487	0.143	0.106	5	<0.0001	1.7
Chr21	rs41378546, rs41451052	CLDN14, UNKNOWN	0.606	0.044	0.627	0.360	5	<0.0001	0.6
	rs41421549, rs7281940	TMPRSS15, UNKNOWN	0.614	0.044	0.215	0.267	5	<0.0001	
Chr22	rs10212068, rs5748617	UNKNOWN, UNKNOWN	0.625	0.077	0.679	0.356	5	<0.0001	0.6
	rs10212068, rs41416344	UNKNOWN, UNKNOWN	0.648	0.068	0.456	0.314	5	<0.0001	
	rs10212068, rs41459445	UNKNOWN, HMGXB4	0.645	0.073	0.541	0.329	5	<0.0001	
	rs10212068, rs1054055	UNKNOWN, C22orf15	0.646	0.070	0.503	0.323	5	<0.0001	
	rs10212068, rs129307	UNKNOWN, UNKNOWN	0.596	0.092	0.789	0.378	5	<0.0001	
	rs10212068, rs41431147	UNKNOWN, TXNRD2	0.616	0.077	0.720	0.363	5	<0.0001	
ChrX	rs1419930, rs41500547	UNKNOWN, DMD	0.666	0.094	0.510	0.314	5	<0.0001	1.8

[a]: Chr chromosome; [b]: false positive rate; [c]: false negative rate; [d]: BMOMDR running time; time unit: *hour* (h)

chance occurrences could be significantly reduced when the CCR value exceeded 0.5, and the maximum value indicated the strongest SNP–SNP interaction [39]. High NMI shows the high level of significance of a SNP–SNP interaction between two SNPs [32]. In all detected SNP–SNP interactions, the maximum *bCCR* was 0.959, and the minimum was 0.579 (mean $\pm$ standard deviation (SD) $= 0.755 \pm 0.097$). The maximum and minimum *bNMI*s were 0.770 and 0.044, respectively, and the mean *bNMI* was 0.269 ± 0.182 (SD). The SNP–SNP interaction rs16926425 and rs7299571 in chromosome 12 revealed the highest *bCCR* (0.959) and *bNMI* (0.770), and FPR and FNR were 0.011 and 0.071, respectively. The gene *SOX5* is RNA-seq in normal kidney tissue [40]. As shown in Table 3, the ten SNP–SNP interactions revealed high values of CCR (>0.8), NMI (>0.4) and Pareto CVC (= 5), and strong significance (p <0.0001). Gene *CACNG1* is found in human adrenal tissue (16 and 10 weeks respectively). *CACNG1* could be the non-additive effect on disease susceptibility in CAD [41] and associate to the tissue-specific circular RNA induction [42]. In SNP–SNP interaction pairs *LOC105376942* and *PLCL2*, *LOC105376942* is RNA-seq found in normal kidney tissue [40], and PLCL2 could play inflammation role to CAD [43]. These SNP pairs may be related to the interactions of CAD. However, further study of the genetic polymorphisms as well as their functional relevance may yield information critical for CAD etiology. The chromosome running times are provided in Table 3. For the average running time of 23 chromosomes, BMOMDR ran for approximately 6.15 hours.

IV. DISCUSSION

In this study, we adopted balancing approaches to modify three steps in MDR and MOMDR, namely generating

K-fold CV datasets, calculating the ratio between cases and controls, and evaluating measures. The results demonstrated that balancing approaches improved the detection success rates of MDR and MOMDR in imbalanced case–control study. MOMDR did not achieve satisfactory detection success rates in imbalanced case–control study (Figs. 2 and 3). CCR measure is widely known as accuracy and is not suitable for MDR in imbalanced case–control study [30]. Yang *et al.* reported the problems associated with imbalance between cases and controls in MDR [31]. The difference between cases and controls can cause most predictive rules to be classified into low-risk groups when the number of cases is less than the number of controls. In this study, both MDR and MOMDR exhibited the same fault in calculating the ratio between cases and controls. Although MOMDR did improve the detection success ratios, both CCR and NMI measures still exhibited abnormal values due to the very high *TN* value in the two-way contingency table. Thus, CCR and NMI measures could have high values in a large number of m-locus combinations, and these high values caused the low detection success rates of MDR and MOMDR for the simulated case–control datasets. Moreover, the results revealed a decrease in the detection success rates along with an increased ratio between cases and controls for MDR and MOMDR. This resulted in decreasing detection success rates along with an increasing *TN* value. The balanced CCR and NMI measures can effectively overcome the problem of a high *TN* value, because *TP*, *FP*, *FN*, and *TN* are transformed to percentages (i.e., *bTP*, *bFP*, *bFN*, *bTN*). Velez *et al.* suggested that percentages of *TP*, *FP*, *FN*, and *TN* can improve the CCR measure [30], and their results corroborated this improvement. Moreover, the balanced NMI exhibited a satisfactory improvement in the detection success ratios (Figs. 2 and 3). We determined that the distribution of cases and controls in each fold CV dataset can also be slightly influenced by a difference between cases and controls. A balanced approach to generate K-fold CV datasets can ensure that a small group can be assigned into each fold CV dataset in imbalanced case–control datasets. Thus, the training models in the fold CV can evaluate sufficient samples in a small group. These three improvements enable BMDR and BMOMDR to be effectively applied to detect SNP–SNP interactions in imbalanced case–control study.

Regarding implementation efficiency, BMDR is similar to MDR, and BMOMDR is similar to MOMDR. For 100 datasets comprising 1000 SNPs with 1600 samples (400 cases and 1200 controls), BMDR was determined to spend on average of 117.4 s to run a complete process on an Intel Core i7 3.60 GHz CPU with 32 GB memory, whereas MDR spent on average 117.3 s. Both MOMDR and BMOMDR required an average of 131.2 s to run a complete process. To determine the optimal m-locus combination in n SNPs using k-fold CV, both MDR and BMDR would require a total computational time of $k\times$ (n choose m)$\times$ the total number of samples $\times$ 3^m times; whereas MOMDR and BMOMDR would require a total computational time of the number of candidates in the k Pareto set $\times k\times$ (n choose m)$\times$ the total number of samples $\times$ 3^m times, in which the number of candidates in the k Pareto set was an average of 3.7 in all the tests.

We combined the balanced function to determine the low-risk and high-risk groups and balanced measures (*bCCR* and *bNMI*) to evaluate the SNP–SNP interactions in imbalanced case–control study. The balanced functions enabled adjusting the single measure of MDR and multiple measures of MOMDR to detect the potential SNP–SNP interactions. The results revealed that BMDR and BMOMDR performed a stronger detection efficiency in simulated imbalanced case–control datasets. BMDR and BMOMDR retain the advantages of MDR: first, BMDR and BMOMDR can effectively minimize false-positive results in detecting SNP–SNP interactions in imbalanced case–control datasets. MOMDR uses a stratified random K-fold CV dataset to select the optimal solution based solely on the ability to predict using independent data. Thus, the training model can avoid overfitting while also minimizing false positives. Second, BMDR and BMOMDR can describe the percentages of cases and controls under m-locus combinations associated with high-risk and low-risk disease groups in imbalanced case–control datasets. Third, BMDR and BMOMDR are model-free methods that do not need a specific inheritance model [15]. Human epistasis (i.e., SNP–SNP interaction) is both chaotic as well as irreducible; human physiology exhibits gradual changes of unknown genetic patterns. The model-free approach is crucial for detecting the SNP–SNP interactions, because simple mono- or oligogenetic traits may be related to epistasis. Fourth, BMDR and BMOMDR are nonparametric methods; thus they perform well with small samples in imbalanced case–control study. Nonparametric statistical analysis methods do not require assumptions regarding the nature of data distributions; this prevents problems related to the use of parametric statistics to detect SNP–SNP interactions [15].

V. CONCLUSION

In this study, we demonstrated balancing approaches that can improve the detection success rates in imbalanced case–control datasets, especially, for substantial differences between cases and controls. A performance assessment of simulated imbalanced datasets revealed that balancing approaches successfully enable MDR and MOMDR to detect SNP–SNP interactions in imbalanced case–control study.

ACKNOWLEDGMENT

This work was partly supported by the Ministry of Science and Technology, China (under Grant 108-2221-E-992-031-MY3, Grant 108-2221-E-214-019-MY3, and Grant 108-2811-E-992-502).

REFERENCES

[1] R. Li, Y. Chen, and J. H. Moore, "Integration of genetic and clinical information to improve imputation of data missing from electronic health records," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 10, pp. 1056–1063, 2019. doi: 10.1093/jamia/ocz041.

[2] L.-Y. Chuang, S.-H. Moi, Y.-D. Lin, and C.-H. Yang, "A comparative analysis of chaotic particle swarm optimizations for detecting single nucleotide polymorphism barcodes," *Artif. Intell. Med.*, vol. 73, pp. 23–33, Oct. 2016.
[3] C.-H. Yang, L.-Y. Chuang, Y.-H. Cheng, Y.-D. Lin, C.-L. Wang, C.-H. Wen, and H.-W. Chang, "Single nucleotide polymorphism barcoding to evaluate oral cancer risk using odds ratio-based genetic algorithms," *Kaohsiung J. Med. Sci.*, vol. 28, no. 7, pp. 362–368, Jul. 2012.
[4] L.-Y. Chuang, Y.-D. Lin, H.-W. Chang, and C.-H. Yang, "An improved PSO algorithm for generating protective SNP barcodes in breast cancer," *PLoS ONE*, vol. 7, no. 5, May 2012, Art. no. e37018.
[5] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, vol. 26, no. 4, pp. 445–455, Feb. 2010.
[6] J.-B. Chen, W.-C. Lee, B.-C. Cheng, S.-H. Moi, C.-H. Yang, and Y.-D. Lin, "Impact of risk factors on functional status in maintenance hemodialysis patients," *Eur. J. Med. Res.*, vol. 22, Dec. 2017, Art. no. 54.
[7] C.-H. Yang, Y.-D. Lin, L.-Y. Chuang, J.-B. Chen, and H.-W. Chang, "Joint analysis of SNP–SNP-environment interactions for chronic dialysis by an improved branch and bound algorithm," *J. Comput. Biol.*, vol. 24, no. 12, pp. 1212–1225, Dec. 2017.
[8] K. V. Steen, "Travelling the world of gene-gene interactions," *Briefings Bioinf.*, vol. 13, no. 1, pp. 1–19, Jan. 2012.
[9] T. F. C. Mackay, "Epistasis and quantitative traits: Using model organisms to study gene–gene interactions," *Nature Rev. Genet.*, vol. 15, no. 1, pp. 22–33, Jan. 2014.
[10] R. J. Urbanowicz, A. S. Andrew, M. R. Karagas, and J. H. Moore, "Role of genetic heterogeneity and epistasis in bladder cancer susceptibility and outcome: A learning classifier system approach," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 4, pp. 603–612, 2013.
[11] T. Hu, Y. Chen, J. W. Kiralis, R. L. Collins, C. Wejse, G. Sirugo, S. M. Williams, and J. H. Moore, "An information-gain approach to detecting three-way epistatic interactions in genetic association studies," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 4, pp. 630–636, 2013.
[12] L. W. Hahn, M. D. Ritchie, and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions," *Bioinformatics*, vol. 19, no. 3, pp. 376–382, Feb. 2003.
[13] X. Zhang, S. P. Huang, F. Zou, and W. Wang, "TEAM: Efficient two-locus epistasis tests in human genome-wide association study," *Bioinformatics*, vol. 26, no. 12, pp. i217–i227, Jun. 2010.
[14] J. Li, J. Dan, C. Li, and R. Wu, "A model-free approach for detecting interactions in genetic association studies," *Briefings Bioinf.*, vol. 15, no. 6, pp. 1057–1068, Nov. 2014.
[15] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *Amer. J. Hum. Genet.*, vol. 69, no. 1, pp. 138–147, Jul. 2001.
[16] C.-H. Yang, Y.-D. Lin, C.-Y. Yen, L.-Y. Chuang, and H.-W. Chang, "A systematic gene–gene and gene–environment interaction analysis of DNA repair genes XRCC1, XRCC2, XRCC3, XRCC4, and oral cancer risk," *OMICS, J. Integr. Biol.*, vol. 19, no. 4, pp. 238–247, Apr. 2015.
[17] C.-H. Yang, Y.-D. Lin, S.-J. Wu, L.-Y. Chuang, and H.-W. Chang, "High order gene-gene interactions in eight single nucleotide polymorphisms of renin-angiotensin system genes for hypertension association study," *BioMed Res. Int.*, vol. 2015, Mar. 2015, Art. no. 454091.
[18] O.-Y. Fu, H.-W. Chang, Y.-D. Lin, L.-Y. Chuang, M.-F. Hou, and C.-H. Yang, "Breast cancer-associated high-order SNP-SNP interaction of CXCL12/CXCR4-related genes by an improved multifactor dimensionality reduction (MDR-ER)," *Oncol. Rep.*, vol. 36, no. 3, pp. 1739–1747, Sep. 2016.
[19] D. Gola, J. M. M. John, K. Van Steen, and I. R. Konig, "A roadmap to multifactor dimensionality reduction methods," *Briefings Bioinf.*, vol. 17, no. 2, pp. 293–308, Mar. 2016.
[20] Y. Chung, S. Y. Lee, R. C. Elston, and T. Park, "Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions," *Bioinformatics*, vol. 23, no. 1, pp. 71–76, Jan. 2007.
[21] C.-H. Yang, L.-Y. Chuang, and Y.-D. Lin, "Multiobjective differential evolution-based multifactor dimensionality reduction for detecting gene–gene interactions," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 12869.
[22] S. Greco, J. R. Figueira, and M. Ehrgott, *Multiple Criteria Decision Analysis*. New York, NY, USA: Springer, 2005.
[23] K. Deb, K. Sindhya, and J. Hakanen, "Multi-objective optimization," *Decision Sciences: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2016, pp. 145–184.
[24] W. Yu, S. Lee, and T. Park, "A unified model based multifactor dimensionality reduction framework for detecting gene–gene interactions," *Bioinformatics*, vol. 32, no. 17, pp. 605–610, Sep. 2016.
[25] F. Ou-Yang, Y. D. Lin, L. Y. Chuang, H. W. Chang, C. H. Yang, and M. F. Hou, "The combinational polymorphisms of ORAI1 gene are associated with preventive models of breast cancer in the Taiwanese," *BioMed Res. Int.*, vol. 2015, Jan. 2015, Art. no. 281263.
[26] C. S. Greene, N. A. Sinnott-Armstrong, D. S. Himmelstein, P. J. Park, J. H. Moore, and B. T. Harris, "Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS," *Bioinformatics*, vol. 26, no. 5, pp. 694–695, Mar. 2010.
[27] C.-H. Yang, L.-Y. Chuang, and Y.-D. Lin, "CMDR based differential evolution identifies the epistatic interaction in genome-wide association studies," *Bioinformatics*, vol. 33, no. 15, pp. 2354–2362, 2017.
[28] J. Gui, J. H. Moore, S. M. Williams, P. Andrews, H. L. Hillege, P. van der Harst, G. Navis, W. H. Van Gilst, F. W. Asselbergs, and D. Gilbert-Diamond, "A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits," *PLoS ONE*, vol. 8, no. 6, Jun. 2013, Art. no. e66545.
[29] S. Lee, M.-S. Kwon, J. M. Oh, and T. Park, "Gene–gene interaction analysis for the survival phenotype based on the Cox model," *Bioinformatics*, vol. 28, no. 18, pp. I582–I588, Sep. 2012.
[30] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genet. Epidemiol.*, vol. 31, no. 4, pp. 306–315, May 2007.
[31] C.-H. Yang, Y.-D. Lin, L.-Y. Chuang, J.-B. Chen, and H.-W. Chang, "MDR-ER: Balancing functions for adjusting the ratio in risk classes and classification errors for imbalanced cases and controls using multifactor-dimensionality reduction," *PLoS ONE*, vol. 8, no. 11, Nov. 2013, Art. no. e79387.
[32] W. S. Bush, T. L. Edwards, S. M. Dudek, B. A. McKinney, and M. D. Ritchie, "Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction," *BMC Bioinf.*, vol. 9, May 2008, Art. no. 238.
[33] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of Database Systems*, L. Liu and M. T. ÖZSU, Eds. New York, NY, USA: Springer, 2016, pp. 1–7.
[34] P. R. Burton and D. G. Clayton, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, Jun. 2007.
[35] M. D. Ritchie, L. W. Hahn, and J. H. Moore, "Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity," *Genetic Epidemiol.*, vol. 24, no. 2, pp. 150–157, Feb. 2003.
[36] J. Namkung, K. Kim, S. Yi, W. Chung, M.-S. Kwon, and T. Park, "New evaluation measures for multifactor dimensionality reduction classifiers in gene–gene interaction analysis," *Bioinformatics*, vol. 25, no. 3, pp. 338–345, Feb. 2009.
[37] R. J. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. M. Fisher, and J. H. Moore, "GAMETES: A fast, direct algorithm for generating pure, strict, epistatic models with random architectures," *Biodata Mining*, vol. 5, Oct. 2012, Art. no. 16.
[38] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. S. Tang, and W. Yu, "Predictive rule inference for epistatic interaction detection in genome-wide association studies," *Bioinformatics*, vol. 26, no. 1, pp. 30–37, Jan. 2010.
[39] C. S. Coffey, P. R. Hebert, M. D. Ritchie, H. M. Krumholz, J. M. Gaziano, P. M. Ridker, N. J. Brown, D. E. Vaughan, and J. H. Moore, "An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene Interactions on risk of myocardial infarction: The importance of model validation," *BMC Bioinf.*, vol. 5, Apr. 2004, Art. no. 49.
[40] L. Fagerberg *et al.*, "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics," *Mol. Cellular Proteomics*, vol. 13, no. 2, pp. 397–406, Feb. 2014.
[41] A. Torkamani, E. J. Topol, and N. J. Schork, "Pathway analysis of seven common diseases assessed by genome-wide association," *Genomics*, vol. 92, no. 5, pp. 265–272, Nov. 2008.
[42] L. Szabo, R. Morey, N. J. Palpant, P. L. Wang, N. Afari, C. Jiang, M. M. Parast, C. E. Murry, L. C. Laurent, and J. Salzman, "Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development," *Genome Biol.*, vol. 17, Jun. 2016, Art. no. 126.
[43] K. Ozaki and T. Tanaka, "Molecular genetics of coronary artery disease," *J. Hum. Genet.*, vol. 61, no. 1, pp. 71–77, Jan. 2016.

CHENG-HONG YANG (M'00–SM'03) received the M.S. and Ph.D. degrees in computer engineering from North Dakota State University, in 1988 and 1992, respectively. He is currently a Chair Professor with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan. He has authored/coauthored over 380 refereed publications and a number of book chapters. His main research areas include evolutionary computation, optimization, bioinformatics, data analysis, and their applications. He is a Fellow of the Institution of Engineering and Technology and the American Biographical Institute. He is an Editorial Board Member of other international journals.

LI-YEH CHUANG received the M.S. degree from the Department of Chemistry, University of North Carolina, in 1989, and the Ph.D. degree from the Department of Biochemistry, North Dakota State University, in 1994. She is currently a Professor with the Department of Chemical Engineering, Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan. She has authored/coauthored over 300 refereed publications. Her main research areas include bioinformatics, biochemistry, and genetic engineering.

YU-DA LIN (M'17) received the M.S. and Ph.D. degrees from the Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Taiwan, in 2011 and 2015, respectively, where he is currently a Research Fellow. He is a Software Engineer and an Adjunct Assistant Professor with the National Kaohsiung University of Science and Technology. He has authored/coauthored over 80 refereed publications. His main research interests include artificial intelligence, biomedical informatics, bioinformatics, and computational biology. He is a member of the IEEE Tainan Section, IEEE Young Professionals, and the IEEE Computational Intelligence Society Membership.