Breast cancer-associated high-order SNP-SNP interaction of *CXCL12/CXCR4*-related genes by an improved multifactor dimensionality reduction (MDR-ER)

OU-YANG FU[1-4], HSUEH-WEI CHANG[2,7,8], YU-DA LIN[5], LI-YEH CHUANG[6], MING-FENG HOU[2,4,9,10] and CHENG-HONG YANG[5]

[1]Graduate Institute of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung 80708; [2]Cancer Center, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung 80708; [3]Department of Surgery, Kaohsiung Municipal Ta-Tung Hospital, Kaohsiung 80145; [4]Department of Surgery, Kaohsiung Medical University Hospital, Kaohsiung 80708; [5]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung 80778; [6]Department of Chemical Engineering and Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung 84001; [7]Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung 80708; [8]Institute of Medical Science and Technology, National Sun Yat-sen University, Kaohsiung 80424; [9]Institute of Clinical Medicine, Kaohsiung Medical University, Kaohsiung; [10]Kaohsiung Municipal Hsiao-Kang Hospital, Kaohsiung 812, Taiwan, R.O.C.

Received February 12, 2016; Accepted March 3, 2016

DOI: 10.3892/or.2016.4956

Abstract. In association studies, the combined effects of single nucleotide polymorphism (SNP)-SNP interactions and the problem of imbalanced data between cases and controls are frequently ignored. In the present study, we used an improved multifactor dimensionality reduction (MDR) approach namely MDR-ER to detect the high order SNP-SNP interaction in an imbalanced breast cancer data set containing seven SNPs of chemokine *CXCL12/CXCR4* pathway genes. Most individual SNPs were not significantly associated with breast cancer. After MDR-ER analysis, six significant SNP-SNP interaction models with seven genes (highest cross-validation consistency, 10; classification error rates, 41.3-21.0; and prediction error rates, 47.4-55.3) were identified. *CD4* and *VEGFA* genes were associated in a 2-loci interaction model (classification error rate, 41.3; prediction error rate, 47.5; odds ratio (OR), 2.069; 95% bootstrap CI, 1.40-2.90; P=1.71E-04) and it also appeared in all the best 2-7-loci models. When the loci number increased, the classification error rates and P-values decreased. The powers in 2-7-loci in all models were >0.9. The minimum classification error rate of the MDR-ER-generated model was shown with the 7-loci interaction model (classification error rate, 21.0; OR=15.282; 95% bootstrap CI, 9.54-23.87; P=4.03E-31). In the epistasis network analysis, the overall effect with breast cancer susceptibility was identified and the SNP order of impact on breast cancer was identified as follows: *CD4* = *VEGFA* > *KITLG* > *CXCL12* > *CCR7* = *MMP2* > *CXCR4*. In conclusion, the MDR-ER can effectively and correctly identify the best SNP-SNP interaction models in an imbalanced data set for breast cancer cases.

Correspondence to: Dr Ming-Feng Hou, Cancer Center, Kaohsiung Medical University Hospital, Kaohsiung Medical University, 100 Shih-Chuan 1st Road, Kaohsiung 80708, Taiwan, R.O.C.
E-mail: mifeho@kmu.edu.tw

Dr Cheng-Hong Yang, Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, 415 Chien Kung Road, Kaohsiung 80778, Taiwan, R.O.C.
E-mail: chyang@cc.kuas.edu.tw

Key words: breast cancer, single nucleotide polymorphism, gene-gene interactions, imbalanced data set, multifactor dimensionality reduction

Introduction

Breast cancer is the most commonly occurring malignant disease in women. The effective identification and screening of biomarkers for breast cancer prediction may reduce the occurrence of breast cancer (1). Single nucleotide polymorphisms (SNPs), the most abundant variants of the human genome (2), have become significant biomarkers for personalized medicine and the recognition of disease/cancer susceptibility (3-5).

Chemokines play a vital role in carcinogenesis (6,7). The representative chemokines are the chemokine ligand CXCL12 also known as stromal cell-derived factor-1 (SDF-1) and its receptor CXC chemokine receptor 4 (CXCR4) which are two main cross-talking factors in tumor microenvironments such as breast carcinogenesis (8). Moreover, the evidence of possible crosstalk between CXCL12, vascular endothelial growth factor (VEGF) (9), matrix metalloproteinase 2 (MMP2) (10), soluble KIT ligands (KITLG) (11), CXCR4, T-cell antigen T4/LEU3 (CD4) (12), and CC-chemokine

receptor 7 (CCR7) (13) are supported by the prediction of protein-protein interaction software (STRING 9) (14) and previously discussed (15). Therefore, the relationship between SNPs for these *CXCL12/CXCR4*-related genes warrants investigation.

Although many disease/cancer-associated SNPs have been reported, the impact of more rare or non-significant SNPs is underestimated and this may partly contribute to 'missing heritability' (16). Lack of a detailed inspection of gene-gene (SNP-SNP) interactions is one of the most common reasons of 'missing heritability' effects (17). For example, several SNPs were reported to be unassociated with diseases and cancers without considering the SNP-SNP interaction (18-20). Recently, the analyses of gene-gene interactions in association studies have been developed in terms of SNP-SNP interactions (21-30). Alternatively, a multifactor dimensionality reduction (MDR) is a data mining technique to provide a non-linear model associated with disease (31). It introduces a strategy with data reduction to identify both high- and low-disease risk associated multiple SNP genotype combinations for SNP-SNP interactions. Biologically meaningful results can be detected without a big sample population because MDR uses a *k*-fold cross-validation (CV). The CV can be regarded as a replication data set technique. The advantage of CV over the repeated random sub-sampling is that all samples are used for training and each sample is used only once for validation. Moreover, MDR has been successfully applied to identify SNP-SNP interactions in various diseases, e.g., atrial fibrillation (32) and coronary artery disease (33). However, MDR cannot quantitatively evaluate the disease susceptibility of genotype combinations (34). Recently, improved MDR methods such as MDR-ER (35) and weighted risk score-based MDR (34) were developed to solve this problem. Moreover, MDR-ER introduces two functions to improve the classification step and an evaluation of error rate in MDR, and it can be applied to the data set of imbalanced numbers of cases and controls.

In our previous investigation (15), we focused on the determination of SNP genotypes of seven SNPs of *CXCL12*-related genes in terms of PCR-restriction fragment length polymorphism (RFLP) analysis for cases and controls. The SNP barcode method was used to investigate the association of potentially combined SNP genotypes of *CXCL12*-related genes. A significant number of genotype combinations of the selected SNPs were reported to be protective against breast cancer and a low risk population of breast cancer patients with these specific combinations was identified. However, their possible SNP-SNP interactions were less investigated (15). This would be important especially for the understanding of a multifactorial interaction in the risk assessment of breast cancer.

In this study, we used the MDR-ER strategy to identify the best breast cancer-associated model for seven SNPs in SNP-SNP interaction of *CXCL12/CXCR4*-related genes, including *CXCL12* (rs1801157, G801A), *CXCR4* (rs2228014, I142I), *VEGF* (rs3025039, C936T, 3'-untranslated region), *KITLG* (rs10506957, intron 1), *MMP2* (rs2287074), *CD4* (rs12812942, intron 3), and *CCR7* (rs3136685, intron 1) which have been reported in several non-MDR disease association studies (15,36-38). The results show that MDR-ER may effectively identify the significant SNP-SNP interaction models of breast cancer susceptibility for imbalanced data sets.

Materials and methods

MDR. MDR is a model-free and non-parametric method for the detection of complex disease/cancer-associated gene-gene (SNP-SNP) interactions (31). The principle of MDR is accomplished by classification to reduce more attributes (loci) into a single attribute (locus). Thus, high-order SNP-SNP interactions can be detected statistically where the data space is transformed into a two-way contingency table.

The MDR procedure is illustrated in Fig. 1. *N* SNPs are considered in a case-control data set, and *M* is the maximum order of SNP-SNP interactions we want to explore, i.e., $M \leq N$. Let *m* be the number of order SNP-SNP interactions ($m \leq M$). The procedures to perform the MDR for detecting the best m-way SNP-SNP interaction model are explained in two parts as follows:

1. Run *k*-fold (usually $k=10$) cross-validation (CV) to detect the best *m*-way interaction (SNP-SNP interaction) model. For each CV fold, steps 1-7 are repeated successively:

Step 1. Select a i^{th} part data set for the test data set and the remaining as the training data set.

Step 2. The pool of all SNPs consists of a set of *m* SNPs.

Step 3. *m* SNPs and their possible multifactor cells are presented in *m*-dimensional space (label 3, Fig. 1):

Equation 1 shows the possible multifactor cells in the *m*-dimensional space:

$$L = \{l_1, l_2, l_3, ..., l_m\} \quad (1)$$

The value of *m* is dependent on the number of considerable factors. Subsequently, a set of *m* genetic and/or environmental factors is chosen.

Step 4. Each multifactor cell is labelled as high-risk when the case/control ratio is higher than or equal to the threshold T (T=1), otherwise the cell is labelled as a low-risk.

The total number of cases and controls are respectively counted in the multifactor cell, and the case/control ratio is calculated by Equation 2.

$$f(L) = \frac{\sum_{j=1}^{P^*} u(L, P_j)}{\sum_{j=1}^{N^*} u(L, N_j)} \quad (2)$$

where

$$u(L, A) = \begin{cases} 1 & \forall l \in A \\ 0 & \forall l \notin A \end{cases}, \forall l \in L$$

where *P* is the case data set; *N*, the control data set; P^*, the number of case groups in a training set; N^*, the number of control groups in a training set; *L*, a vector of variable combinations. The function *u()* determines a score of '1' if all elements *l* in *L* match the cases or controls; otherwise given a score of '0'.

The high/low risk in each multifactor cell is determined. Each multifactor cell is labelled as 'H' or 'L' symbol. Label 'H' indicates that ratio in the multifactor cell meets or exceeds a threshold (high-risk group); otherwise, the label is 'L' (low-risk group). The threshold is equal to the one in a balanced data set.

Step 5. Repeat steps 2-5 until all possible sets of *m* SNPs are evaluated.

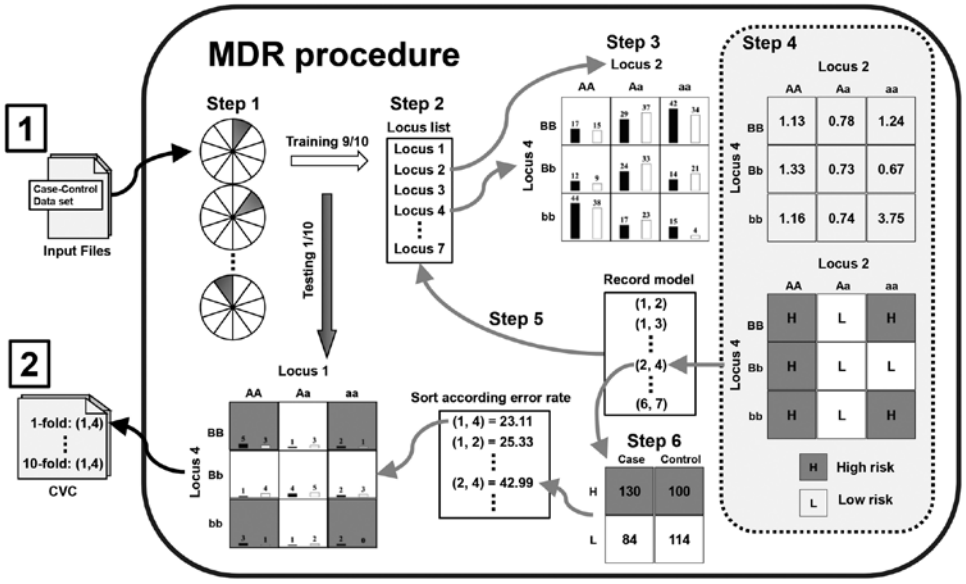

Figure 1. MDR flowchart.

Step 6. Evaluate all possible combinations of m SNPs, build $\binom{N}{m}$ contingency tables, and correspondingly get $\binom{N}{m}$ training error rates. The model with the minimum training error rate (classification error rate) is selected in each CV, and the prediction error rate of the model is evaluated using the independent test data set.

Step 4 reduces the possible combinations in n-loci into a two-way contingency table [true positive (TP), false positive (FP), true negative (TN), and false negative (FN)]. *TP* is the total number of labeled 'H' in the case data. *FP* is the total number of labeled 'H' in the control data. *FN* is the total number of labeled 'L' in the case data. *TN* is the total number of labeled 'L' in the control data. Thus, statistical analysis can be used to evaluate the n-loci effect. Equation 3 is used to evaluate model classification error rate and prediction error rate.

$$f(C) = \frac{FN + FP}{TP + FN + FP + TN} \quad (3)$$

where C is the possible combination of m SNPs.

Step 7. Repeat steps 1-7 until k=10.

2. Collect CV to create cross-validation consistency (CVC) and select the highest frequency with CVC as the best SNP-SNP interaction models. The classification error rate of the best model is calculated as the averaged classification error rates in those CVC included k models, and the prediction error rate is calculated as the averaged prediction error rates in those CVC included k models.

After classification error rate evaluates all the possible SNP-SNP interaction models, the model with minimum error rate is regarded as the best SNP-SNP interaction model of training data at i^{th}-fold CV. This best SNP-SNP interaction model is then evaluated by testing the data for evaluating the prediction error rate. Thus, the aforementioned steps 1-7 are repeated in each fold CV. The best SNP-SNP interaction model with a minimum classification error rate is chosen at each CV, and the ten best SNP-SNP interaction models of ten-fold CV are classified by the same model. The number of best models in a classified group are counted and named CVC. Finally, a nSNP-SNP interaction model of CVC with highest occurrence frequency results is regarded as the best model. If the equal frequency of CVC occurs in two or more models, then the model found first is the best SNP-SNP interaction model. The classification error rate of the final selected best model is calculated from averaged k classification error rates of models included in highest frequency with CVC.

Improved MDR for an imbalanced data set (MDR-ER). MDR-ER was proposed previously (35). MDR-ER introduced the percentage concept to improve Equations 2 and 3 for imbalanced data. Equation 2 of MDR is modified as Equation 4 which evaluates the ratio of the percentages of cases and controls.

$$f(L) = \frac{N^* [\sum_{j=1}^{P^*} u(L, P_j)]}{P^* [\sum_{j=1}^{N^*} u(L, N_j)]} \quad (4)$$

where

$$u(L, A) = \begin{cases} 1 & \forall l \in A \\ 0 & \forall l \notin A \end{cases}, \forall l \in L$$

where P is the case data set; N, the control data set; P^*, the number of case groups in a training set; N^*, the number of control groups in a training set; L, a vector of variable combinations.

The function $u()$ determines a score of '1' if all elements l in L match the cases or controls; otherwise it is given a score of '0'. As in the classification of MDR, label 'H' is the ratio in the multifactor cell meeting or exceeding a threshold (=1); otherwise the label is 'L'.

MDR-ER modifies the classification/prediction error rate function of MDR (Equation 3). Equation 5 is the adjusted classification/prediction error rate function based

Table I. The performance of eight individual SNPs for the case and the control groups.

Locus	Genotypes	Cases (n=220) n (%)	Controls (n=334) n (%)	Odds ratio[a]	Bootstrap odds ratio[b]	Bootstrap 95% CI[c]
CD4 gene rs12812942	AA	128 (58.2)	174 (52.1)			
	AT	76 (34.5)	141 (42.2)	0.733	0.746	(0.521, 1.064)
	TT	16 (7.3)	19 (5.7)	1.145	1.228	(0.540, 2.361)
CCR7 gene rs3136685	GG	77 (35.0)	107 (32.0)			
	GA	114 (51.8)	180 (53.9)	0.880	0.897	(0.590, 1.293)
	AA	29 (13.2)	47 (14.1)	0.857	0.887	(0.474, 1.464)
CXCR4 gene rs2228014	CC	151 (68.6)	254 (76.0)			
	CT	63 (28.6)	73 (21.9)	1.452	1.451	(0.863, 2.137)
	TT	6 (2.7)	7 (2.1)	1.442	1.790	(0.369, 5.641)
CXCL12 gene rs1801157	GG	106 (48.2)	175 (52.4)			
	GA	98 (44.5)	136 (40.7)	1.190	1.209	(0.818, 1.669)
	AA	16 (7.3)	23 (6.9)	1.148	1.198	(0.534, 2.216)
VEGFA gene rs3025039	CC	155 (70.5)	211 (63.2)			
	CT	59 (26.8)	117 (35.0)	0.686	0.680	(0.438, 0.976)
	TT	6 (2.7)	6 (1.8)	1.361	1.626	(0.340, 4.889)
MMP2 gene rs2287074	GG	113 (51.4)	164 (49.1)			
	GA	93 (42.3)	139 (41.6)	0.971	0.976	(0.677, 1.392)
	AA	14 (6.4)	31 (9.3)	0.655	0.676	(0.293, 1.278)
KITLG gene rs10506957	TT	133 (60.5)	182 (54.5)			
	TC	69 (31.4)	133 (39.8)	0.710	0.721	(0.475, 1.037)
	CC	18 (8.2)	19 (5.7)	1.296	1.390	(0.608, 2.645)

[a]Data from our previous results (15). [b]Bootstrap odds ratio: odds ratio is adjusted by deviation which is calculated by bootstrap results based on 1,000 bootstrap samples. [c]Bootstrap results are based on 1,000 bootstrap samples.

on the arithmetic mean of sensitivity and specificity. The adjusted classification/prediction error rate is algebraically identical to the error rate if the case and the control data sets are imbalanced.

$$f(C) = 0.5 \times \left(\frac{FN}{TP + FN} + \frac{FP}{FP + TN} \right) \quad (5)$$

where *TP* is the total number of labeled 'H' in the case data; *FP*, the total number of labeled 'H' in the control data; *FN*, the total number of labeled 'L' in the case data; *TN*, the total number of labeled 'L' in the control data.

Statistical analyses. Statistical analyses were evaluated by the two-way contingency table and its TP, FP, TN, and FN values were respective averages of TP, FP, TN, and FN in that CVC included the best SNP-SNP interaction models in the training models. The disease risks of SNPs were evaluated by SPSS version 19.0 (SPSS, Inc., Chicago, IL, USA), including the odds ratio (OR) or the bootstrap OR and its 95% confidence interval (CI). P-values were used to define significant differences between the cases and the controls. The Power and Sample Size Calculations (PS) tool (39) was used to evaluate the power with statistical analysis. The Power can determine the null hypothesis (OR=1) with Type I error probability α (=0.05), i.e., the probability of rejecting the OR=1.

Results

Data set. A breast cancer data set containing breast cancer patients (n=220) and normal controls (n=334) was obtained from our previous study (15). The genotype information is available at http://bioinfo.kmu.edu.tw/brca-7SNP_all_BPSO.xls. Seven SNPs of *CXCL12*-related genes were included, such as *CD4* (rs12812942), *CCR7* (rs3136685), *CXCR4* (rs2228014), *CXCL12* (rs1801157), *VEGFA* (rs3025039), *MMP2* (rs2287074), and *KITLG* (rs10506957). However, the data mining strategy-based SNP-SNP interaction was not addressed. In the present study, we applied the MDR-ER to detect the best SNP-SNP interaction model with a significant difference between breast cancer (cases) and normal (controls) groups.

Single SNP analysis. Table I shows the OR, bootstrap OR and its 95% CI of each single SNP in the breast cancer association. The distribution of genotypes for most SNPs showed no significant difference between the case and the control groups.

SNP-SNP interaction analysis - the determination of the best model in CVC. Although many 2-loci models exist, only the significant 2-loci SNP-SNP interaction models are provided in Table II. The best model in CVC was defined by the model which has the minimum classification error rate.

Table II. 2-Loci SNP-SNP interactions among seven SNPs assessed by MDR-ER[a].

2-Loci	Odds ratio (95% CI)	Bootstrap 95% CI	P-value	Error rate
CD4 + CCR7	1.519 (1.04, 2.23)	(1.04, 2.25)	0.032	0.454
CD4 + CXCR4	1.549 (1.06, 2.26)	(1.07, 2.27)	0.022	0.449
CD4 + VEGFA	2.069 (1.44, 2.98)	(1.40, 2.90)	1.71E-04	**0.413**
CD4 + KITLG	2.000 (1.28, 3.14)	(1.24, 3.24)	0.002	0.441
CCR7 + VEGFA	1.537 (1.06, 2.24)	(1.06, 2.27)	0.025	0.450
CCR7 + KITLG	1.491 (1.04, 2.14)	(1.04, 2.25)	0.027	0.452
CXCR4 + CXCL12	1.520 (1.05, 2.21)	(1.05, 2.17)	0.005	0.451
CXCR4 + VEGFA	1.714 (1.14, 2.57)	(1.03, 2.54)	0.009	0.448
CXCR4 + KITLG	1.781 (1.19, 2.67)	(1.19, 2.74)	0.005	0.440
CXCL12 + KITLG	1.506 (1.05, 2.17)	(1.04, 2.18)	0.027	0.449
VEGFA + MMP2	1.537 (1.05, 2.24)	(1.07, 2.23)	0.026	0.450
VEGFA + KITLG	1.732 (1.21, 2.49)	(1.19, 2.44)	0.003	0.432
MMP2 + KITLG	1.808 (1.25, 2.62)	(1.28, 2.67)	0.002	0.430

[a]All 2-loci SNP-SNP interactions with significant testing accuracy were identified by the MDR-ER method. The minimum error rate is marked in bold type.

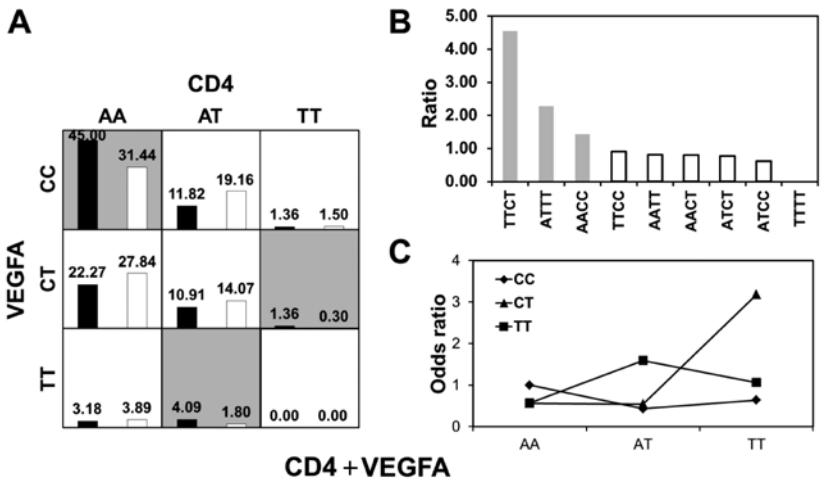

Figure 2. The best 2-loci interactions are analyzed by MDR-ER. The SNP-SNP interaction between *CD4* and *VEGFA* genes was used as an example. (A) The joint distribution of the genotypes of the *CD4* and *VEGFA* genes. High- and low-risks are expressed under the grey and white backgrounds, respectively. The left bar (black color) in a cell represents the frequency of counting cases and the right bar (white color) represents the frequency of counting controls. (B) The ratio between the cases and the controls in each cell. The gray and white bars represent the high- and low-risks, respectively. (C) The deviation with pattern of the odds ratio from two-way interaction models. The reference for the odds ratio is the combination of the major genotypes of these two genes.

Among 2-loci SNP-SNP interaction models, the minimum classification error of the best 2-loci model (*CD4 + VEGFA*) in CVC is 0.413. Fig. 2 shows the details of the 2-loci interaction (*CD4 + VEGFA*) based on a MDR-ER method in a breast cancer study. The percentages of genotypes for the *CD4* and the *VEGFA* genes in the cases and the controls are shown (Fig. 2A). Cells with a SNP combination of higher risk of breast cancer are marked with a gray background, i.e., the percentages of the cases were higher than those of the controls. Fig. 2B provides the ratios of each 2-loci SNP-SNP interaction. SNP combinations of breast cancer risk (OR>1) are identified (marked in gray colors). Fig. 2C shows that genotype effects (CC, CT, and TT) of the *VEGFA* gene were not additive to the genotypes (AA, AT, and TT) of the *CD4* gene - it indicates an interaction between these two genes. Similarly, all the best models in CVC in terms of 3-7-loci SNP-SNP interaction models were selected (see details in Table III). In the 4-loci SNP-SNP interaction, the interactions (*CD4 + CCR7 + VEGFA + KITLG*), (*CD4 + VEGFA + MMP2 + KITLG*), and (*CD4 + CXCL12 + VEGFA + KITLG*) had an equal frequency in CVC, indicating that they have the same importance in breast cancer. Table III shows the first identified model, i.e., *CD4 + CXCL12 + VEGFA + KITLG*.

SNP-SNP interaction analysis - error rates. Table III shows the best 2-7-loci SNP-SNP interaction models using MDR-ER analysis. When the loci number was increased, the best SNP-SNP interaction training models showed higher consistency of breast cancer and lower classification error rates (ranging from 41.3 to 21.0). The prediction error rates of the best SNP-SNP

Table III. Analysis results of the best 2- to 7-loci SNP-SNP interaction models using MDR-ER.

Loci number (gene combination)	Consistency	Classification error (%)	Prediction error (%)	Power	OR (95% CI)	Bootstrap 95% CI	P-value
2-Loci (*CD4*, *VEGFA*)	8/10	41.3	47.5	0.978	2.069 (1.44, 2.98)	(1.40, 2.90)	1.71E-04
3-Loci (*CD4*, *VEGFA*, *KITLG*)	8/10	39.1	47.4	1.000	2.652 (1.81, 3.90)	(1.75, 3.77)	1.58E-06
4-Loci (*CD4*, *CXCL12*, *VEGFA*, *KITLG*)	2/10	35.8	52.5	1.000	3.318 (2.27, 4.86)	(2.22, 4.74)	1.36E-09
5-Loci (*CD4*, *CCR7*, *VEGFA*, *MMP2*, *KITLG*)	8/10	31.1	47.8	1.000	5.008 (3.38, 7.42)	(3.40, 7.46)	8.47E-16
6-Loci (*CD4*, *CCR7*, *CXCL12*, *VEGFA*, *MMP2*, *KITLG*)	9/10	25.6	52.2	1.000	8.900 (5.82, 13.61)	(5.69, 13.29)	1.54E-23
7-Loci (*CD4*, *CCR7*, *CXCR4*, *CXCL12*, *VEGFA*, *MMP2*, *KITLG*)	10/10	21.0	55.3	1.000	15.282 (9.67, 24.14)	(9.54, 23.87)	4.03E-31

[a]Bootstrap results are based on 1,000 bootstrap samples.

interaction training models in 2-7-loci are the region between 47.4 and 55.3. The 7-loci model shows a minimum classification error rate of 21.0% and the 3-loci model shows a minimum prediction error rate of 47.4, which are observed by chance in randomized data based on the null hypothesis of no association. In addition, the frequency differences between the case and the control groups of the best MDR-ER-generated 2-7-loci models were significant (P<0.01). The 2-7-loci models suggest that all SNPs of *CD4*, *CCR7*, *CXCR4*, *CXCL12*, *VEGFA*, *MMP2*, and *KITLG* genes are associated with breast cancer.

SNP-SNP interaction analysis - OR, P-value, and power analysis. In Table III, the OR values in the 2-7-loci models were increased from 2.069 to 15.282 and the 95% CI was 1.44-24.14. The bootstrapping in 1,000 samples with the adjustments of 95% CI of OR (95% bootstrap CI) values were adjusted from 1.40 to 23.87 for the 2-7-loci models. The P-values of the 2-7-loci models decreased from 1.71E-04 to 4.03E-31. The P-values were decreased and 95% bootstrap CI values were increased when the loci numbers increased, indicating that the risk for breast cancer was increased by a combined effect of SNPs. In the example of the 7-loci SNP-SNP interaction model, the power analysis in the case data set showed that the probability of exposure among controls was 0.231 (77/334, 77 is *FP*). The powers in the 2-7-loci, ranging from 0.978 to 1.000, showed that occurrence probability in all models was >0.9. These findings suggest that all these seven SNPs are highly associated with breast cancer.

SNP-SNP interaction analysis - SNP-SNP interaction network. Fig. 3 shows an SNP-SNP interaction network of a 7-loci SNP-SNP interaction model associated with breast cancer susceptibility. The epistasis networks were constructed by integrating 13 significant 2-loci SNP-SNP interaction models (Table II) where the non-significant interactions are not shown. The susceptibility to breast cancer can be explained by showing the details of a two-factor interaction based on the MDR-ER method in the example of *CD4* + *VEGFA* (Fig. 2) as well as other two-factor interactions (data not shown). In Fig. 3, genes (SNPs) involved in one or more significant interactions are represented as nodes, and the pairs of genes (SNPs) with significant interactions are connected by lines. Each line is labeled with the corresponding OR value and the thickness of the lines represent the stronger OR values. Thus, Fig. 3 can clearly illustrate how combined effects are associated with SNP (genes) to generate the overall effect.

Discussion

Many breast cancer studies have reported several breast cancer-associated genes, including *CD4* (15), *CCR7* (13,15), *CXCR4* (15), *CXCL12* (36,40), *VEGF* (41), and *MMP2* (10). In the individual SNP analysis (Table I), only the SNP of the *VEGFA* gene was found to be breast cancer-associated but it was still not significant after Bonferroni's correction. In general, these rare and non-significant SNPs are commonly ignored and we propose that they may partly solve the problem of missing heritability. However, epistasis of rare SNPs is not the only way to address missing heritability of common complex traits. Without the help of computation, rare SNPs could hardly explain missing heritability in asthma (42). To improve the sensitivity of association, these hitherto non-significant SNPs were further combined and the possible combined effects associated with breast cancer were examined in this study. Likewise, some SNPs may have additive SNP-SNP interactions or non-additive effects for genomic monitoring and prediction of complex traits (43).

Based on imbalanced cases and controls, the MDR-ER algorithm explored six multiple SNP loci with significant associations with breast cancer in terms of the 2-7-loci with OR values (Table III). If the true OR of the best model for breast cancer in exposed subjects relative to unexposed subjects is

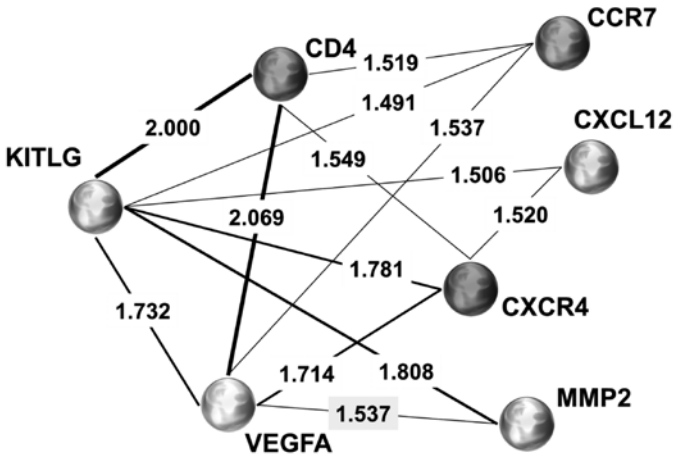

Figure 3. SNP-SNP interaction network. The epistasis networks of 7-loci models for SNP-SNP interaction were found to be associated with breast cancer. Significant SNP-SNP interactions (P<0.05) in these multi-foci models are connected by lines, and the strength of interaction is labeled with OR values. The thicker and thinner lines represent the higher and lower interactions, respectively.

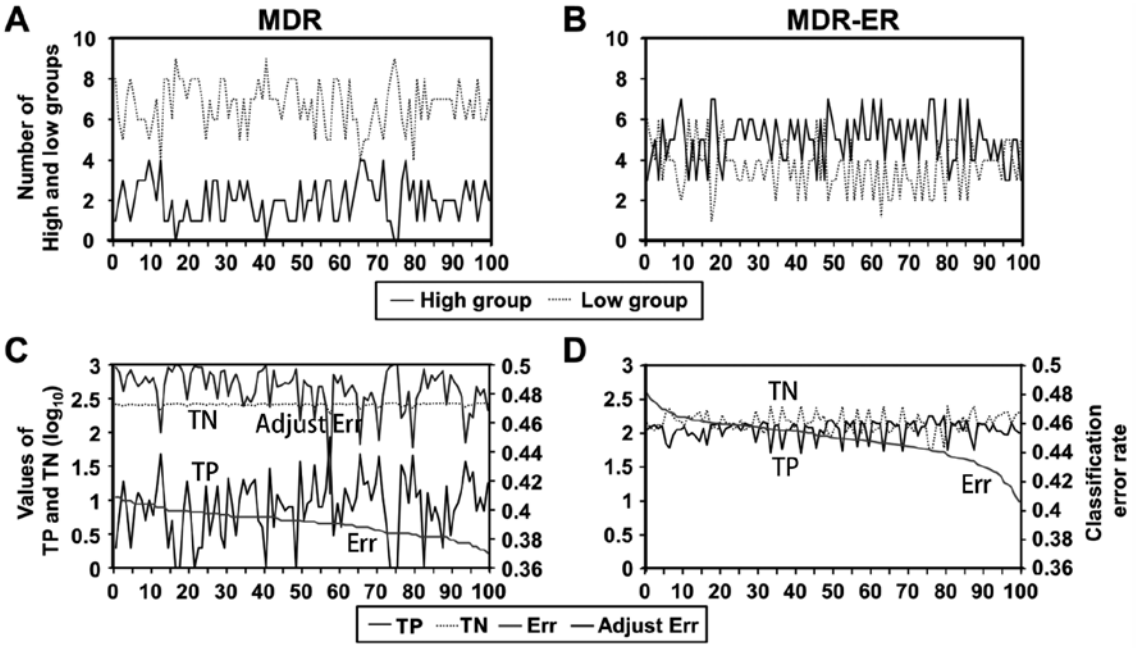

Figure 4. Comparison difference between MDR and MDR-ER in the imbalanced data set using the 2-loci SNP-SNP interaction. (A and B) The numbers of high- and low-risk groups in the algorithm implementation, respectively. The distribution differences between the numbers of high- and low-risk groups are shown. (C and D) The frequencies of TP and TN, and the classification error rate in algorithm implementation, respectively. The left and right scales of the vertical axis show the $\log_{10}$ value for the total numbers of TP and TN, and the classification error rate, respectively. The top solid line in (C) indicates the adjusted error rate based on Equation 5. The horizontal axis indicates the 100 models of the 2-loci SNP-SNP interaction. All models are sorted by the classification error rate and selected by systematic sampling.

15.282 (OR value of 7-loci SNP-SNP interaction model), we can reject the null hypothesis that OR=1 with the probability (power) 1.000. The type I error probability associated with the test of this null hypothesis was 0.05. Furthermore, the OR values in the 2-7-loci showed a gradual increase, suggesting that all genes (SNPs) in the 2-7-loci were additive and highly associated with breast cancer.

The key evaluations in MDR-ER represent a classification error rate and prediction error rate which aim to correctly evaluate the proportion of an incorrect prediction. The incorrect prediction error in MDR-ER is a measurement for internal validation that avoids finding an association by chance in the test sample. When the error rate is <50%, the associations by chance are significantly reduced. Thus, in Table III, the 4-, 6- and 7-loci results (prediction error rates >50%) cannot be considered as the predictive models. While the 2-, 3- and 5-loci results showed prediction error rates ~47%, and the 3-loci had the lowest prediction error rate. Therefore, these MDR-ER generated SNP-SNP interaction models are very effective for classification of the disease risk, and suggest that SNP combination of *CD4*, *VEGFA*, and *KITLG* genes provide the best predictive models. Moreover, a SNP combination of *CD4*

and *VEGFA* genes is included in all the best 2-7 loci models, suggesting that it is really driving the associations in breast cancer.

For example in Fig. 3, *CXCL12* was only significantly associated with *CXCR4* and *KITLG* (OR=1.520 and 1.506, respectively) but it had an overall effect by integrating *KITLG* with *CD4*, *CCR7*, *VEGFA*, and *MMP2*. All lines in Fig. 3 were found to connect to *KITLG*, indicating that all SNPs from all the listed genes were joined together with the SNP of *KITLG* which is the main connector of this network. The breast cancer associated effect of *KITLG* has been less mentioned previously, however, the importance of *KITLG* was detectable using a MDR-ER algorithm-based SNP-SNP interaction.

According to the performance of OR values, the *CD4* and *VEGFA* (2.069) were the best models in the 2-loci and *CD4* and *KITLG* (2.000) were subsequently integrated with this best 2-loci model, where the difference was very small (0.069). They showed a strong interaction in breast cancer, but *CD4* and *VEGFA* were more highly associated with breast cancer than *KITLG*. Thus, an order of SNPs with breast cancer association can be suggested as follows: *CD4* = *VEGFA* > *KITLG* > *CXCL12* > *CCR7* = *MMP2* > *CXCR4*. This combined effect order can detect the overall effect of the impact of a gene (SNP) on breast cancer. Therefore, we suggest that *CD4*, *VEGFA*, *KITLG*, *CXCL12*, *CCR7*, *MMP2*, and *CXCR4* genes have an overall effect with breast cancer susceptibility.

In the 2-loci SNP-SNP interaction model, the comparison between MDR and MDR-ER faced with the imbalanced data set are shown in Fig. 4. When using MDR, the large number of the group (either the cases or the controls) of the imbalanced data set can affect the high- or low-risk classification and the evaluation for classification error rate compared to that of MDR-ER. Using MDR analysis in the present study (220 cases and 334 controls), the number of low-risk groups was greater than the high-risk groups in 100 selected models (Fig. 4A). MDR-based classification error rates were 0.37-0.4 and TPs were always higher than TNs (Fig. 4C). However, the averages of sensitivity and specificity in the 100 models were 0.07 and 0.96, respectively. This result suggests that this low classification error rate and high OR value using MDR are generated by its high TN but low TP values implying a low sensitivity for disease detection.

In contrast, MDR-ER showed balanced frequencies between the numbers of high- and low-risk groups in each model (Fig. 4B). Its TN values were not always higher than its TP values and classification error rates were also improved in MDR-ER (Fig. 4D). The averages of sensitivity and specificity in the 100 models using MDR-ER were 0.589-0.511, respectively.

MDR-ER is designed to combine two improved functions to measure the low- and high-risk groups and it evaluates the classification error to select the best model. Thus, MDR-ER allows for gene-gene interaction detection studies on imbalanced data sets without the balanced study population technologies. Moreover, MDR-ER holds the original MDR characteristics, including a non-parametric method and assumes no particular genetic model. Therefore, it can provide strong detecting ability to imbalanced data sets for analyzing the possible gene-gene and gene-environment interactions. When imbalanced data sets are used, MDR-ER has several advantages which are as follows: i) MDR-ER improves the classification function to effectively classify cells into low- and high-risk groups; thus the number of TPs can be increased; ii) the final best model has a low error rate and a high sensitivity for disease prediction; iii) MDR-ER only adjusts and improves two formulas and therefore the number of procedures and parameters are not increased; and iv) MDR-ER is based on the quantitative value of the ratios representing better classifications results.

The MDR and MDR-ER can be limited by the overall running time due to the rapidly growing total number of SNP combinations when the number of SNPs, order, and sample size are increased. Substantial time requirements may limit the multiple tests in finding more complex interactions between genes related to diseases.

In conclusion, we demonstrated that MDR-ER can effectively and correctly identify the best SNP-SNP interaction models in an imbalanced data set. The joint effect of SNP-SNP interactions of chemokine *CXCL12*/*CXCR4* pathway genes in breast cancer susceptibility was also identified. MDR-ER has potential to apply to many other associated studies with imbalanced data sets.

Acknowledgements

This study was supported by funds of the Ministry of Science and Technology, Taiwan (MOST 102-2221-E-151-024-MY3 and MOST 104-2320-B-037-013-MY3), the National Sun Yat-sen University-KMU Joint Research Project (no. NSYSU-KMU 105-p022), and the Health and Welfare Surcharge of Tobacco Products, the Ministry of Health and Welfare, Taiwan, Republic of China (MOHW105-TDU-B-212-134005). We also thank Dr Hans-Uwe Dahms for his help with the English editing.

References

1. Pepe MS and Janes HE: Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. J Natl Cancer Inst 100: 978-979, 2008.
2. Chuang LY, Yang CH, Tsui KH, Cheng YH, Chang PL, Wen CH and Chang HW: Restriction enzyme mining for SNPs in genomes. Anticancer Res 28 (4A): 2001-2007, 2008.
3. Chang HW, Chuang LY, Tsai MT and Yang CH: The importance of integrating SNP and cheminformatics resources to pharmacogenomics. Curr Drug Metab 13: 991-999, 2012.
4. Yang CH, Cheng YH, Chuang LY and Chang HW: Drug-SNPing: An integrated drug-based, protein interaction-based tagSNP-based pharmacogenomics platform for SNP genotyping. Bioinformatics 29: 758-764, 2013.
5. Liu L, Hua FZ, Cao JQ and Zhu PQ: VEGF 936C>T polymorphism and breast cancer risk: Evidence needed further clarification. Breast Cancer Res Treat 127: 569-571, 2011.
6. Homey B, Müller A and Zlotnik A: Chemokines: Agents for the immunotherapy of cancer? Nat Rev Immunol 2: 175-184, 2002.
7. Zabaleta J, Lin HY, Sierra RA, Hall MC, Clark PE, Sartor OA, Hu JJ and Ochoa AC: Interactions of cytokine gene polymorphisms in prostate cancer risk. Carcinogenesis 29: 573-578, 2008.
8. Luker KE and Luker GD: Functions of CXCL12 and CXCR4 in breast cancer. Cancer Lett 238: 30-41, 2006.
9. Yan Y, Liang H, Li T, Guo S, Li M, Li S and Qin X: Vascular endothelial growth factor +936C/T polymorphism and breast cancer risk: A meta-analysis of 13 case-control studies. Tumour Biol 35: 2687-2692, 2014.
10. Zhou Y, Yu C, Miao X, Tan W, Liang G, Xiong P, Sun T and Lin D: Substantial reduction in risk of breast cancer associated with genetic polymorphisms in the promoters of the matrix metalloproteinase-2 and tissue inhibitor of metalloproteinase-2 genes. Carcinogenesis 25: 399-404, 2004.

11. Jin DK, Shido K, Kopp HG, Petit I, Shmelkov SV, Young LM, Hooper AT, Amano H, Avecilla ST, Heissig B, *et al*: Cytokine-mediated deployment of SDF-1 induces revascularization through recruitment of $CXCR4^+$ hemangiocytes. Nat Med 12: 557-567, 2006.
12. Viehl CT, Frey DM, Phommaly C, Chen T, Fleming TP, Gillanders WE, Eberlein TJ and Goedegebuure PS: Generation of mammaglobin-A-specific CD4 T cells and identification of candidate CD4 epitopes for breast cancer vaccine strategies. Breast Cancer Res Treat 109: 305-314, 2008.
13. Cabioglu N, Yazici MS, Arun B, Broglio KR, Hortobagyi GN, Price JE and Sahin A: CCR7 and CXCR4 as novel biomarkers predicting axillary lymph node metastasis in T1 breast cancer. Clin Cancer Res 11: 5686-5693, 2005.
14. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, *et al*: STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41 (D1): D808-D815, 2013.
15. Lin GT, Tseng HF, Yang CH, Hou MF, Chuang LY, Tai HT, Tai MH, Cheng YH, Wen CH, Liu CS, *et al*: Combinational polymorphisms of seven CXCL12-related genes are protective against breast cancer in Taiwan. OMICS 13: 165-172, 2009.
16. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH and Nadeau JH: Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446-450, 2010.
17. Cordell HJ: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392-404, 2009.
18. Yu KD, Di GH, Fan L, Chen AX, Yang C and Shao ZM: Lack of an association between a functional polymorphism in the interleukin-6 gene promoter and breast cancer risk: A meta-analysis involving 25,703 subjects. Breast Cancer Res Treat 122: 483-488, 2010.
19. Hsieh MH, Chong IW, Hwang JJ, Lee CH, Ho CK, Yu ML, Huang CT, Lee CY, Wu MT and Christiani DC: Lack of associations between several polymorphisms in cytokine genes and the risk of chronic obstructive pulmonary diseases in Taiwan. Kaohsiung J Med Sci 24: 126-137, 2008.
20. Pinheiro UB, de Carvalho Fraga CA, Mendes DC, Marques-Silva L, Farias LC, de Souza MG, Soares MB, Jones KM, Santos SH, de Paula AM, *et al*: p16 (CDKN2A) SNP rs11515 was not associated with head and neck carcinoma. Tumour Biol 35: 6113-6118, 2014.
21. Yang CH, Chuang LY, Chen YJ, Tseng HF and Chang HW: Computational analysis of simulated SNP interactions between 26 growth factor-related genes in a breast cancer association study. OMICS 15: 399-407, 2011.
22. Chen JB, Chuang LY, Lin YD, Liou CW, Lin TK, Lee WC, Cheng BC, Chang HW and Yang CH: Preventive SNP-SNP interactions in the mitochondrial displacement loop (D-loop) from chronic dialysis patients. Mitochondrion 13: 698-704, 2013.
23. Yang CH, Chuang LY, Cheng YH, Lin YD, Wang CL, Wen CH and Chang HW: Single nucleotide polymorphism barcoding to evaluate oral cancer risk using odds ratio-based genetic algorithms. Kaohsiung J Med Sci 28: 362-368, 2012.
24. Chen JB, Chuang LY, Lin YD, Liou CW, Lin TK, Lee WC, Cheng BC, Chang HW and Yang CH: Genetic algorithm-generated SNP barcodes of the mitochondrial D-loop for chronic dialysis susceptibility. Mitochondrial DNA 25: 231-237, 2014.
25. Jeon S, Choi JY, Lee KM, Park SK, Yoo KY, Noh DY, Ahn SH and Kang D: Combined genetic effect of CDK7 and ESR1 polymorphisms on breast cancer. Breast Cancer Res Treat 121: 737-742, 2010.
26. Li J, Humphreys K, Heikkinen T, Aittomäki K, Blomqvist C, Pharoah PD, Dunning AM, Ahmed S, Hooning MJ, Martens JW, *et al*: A combined analysis of genome-wide association studies in breast cancer. Breast Cancer Res Treat 126: 717-727, 2011.
27. Sfar S, Saad H, Mosbah F and Chouchane L: Combined effects of the angiogenic genes polymorphisms on prostate cancer susceptibility and aggressiveness. Mol Biol Rep 36: 37-45, 2009.
28. Cherdyntseva NV, Denisov EV, Litviakov NV, Maksimov VN, Malinovskaya EA, Babyshkina NN, Slonimskaya EM, Voevoda MI and Choinzonov EL: Crosstalk between the FGFR2 and TP53 genes in breast cancer: Data from an association study and epistatic interaction analysis. DNA Cell Biol 31: 306-316, 2012.
29. Choudhury JH, Choudhury B, Kundu S and Ghosh SK: Combined effect of tobacco and DNA repair genes polymorphisms of XRCC1 and XRCC2 influence high risk of head and neck squamous cell carcinoma in northeast Indian population. Med Oncol 31: 67, 2014.
30. Lane HY, Tsai GE and Lin E: Assessing gene-gene interactions in pharmacogenomics. Mol Diagn Ther 16: 15-27, 2012.
31. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF and Moore JH: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69: 138-147, 2001.
32. Asselbergs FW, Moore JH, van den Berg MP, Rimm EB, de Boer RA, Dullaart RP, Navis G and van Gilst WH: A role for CETP TaqIB polymorphism in determining susceptibility to atrial fibrillation: A nested case control study. BMC Med Genet 7: 39, 2006.
33. Agirbasli M, Guney AI, Ozturhan HS, Agirbasli D, Ulucan K, Sevinc D, Kirac D, Ryckman KK and Williams SM: Multifactor dimensionality reduction analysis of MTHFR, PAI-1, ACE, PON1, and eNOS gene polymorphisms in patients with early onset coronary artery disease. Eur J Cardiovasc Prev Rehabil 18: 803-809, 2011.
34. Li CF, Luo FT, Zeng YX and Jia WH: Weighted risk score-based multifactor dimensionality reduction to detect gene-gene interactions in nasopharyngeal carcinoma. Int J Mol Sci 15: 10724-10737, 2014.
35. Yang CH, Lin YD, Chuang LY, Chen JB and Chang HW: 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 8: e79387, 2013.
36. Hassan S, Baccarelli A, Salvucci O and Basik M: Plasma stromal cell-derived factor-1: Host derived marker predictive of distant metastasis in breast cancer. Clin Cancer Res 14: 446-454, 2008.
37. Cheong JY, Cho SW, Choi JY, Lee JA, Kim MH, Lee JE, Hahm KB and Kim JH: RANTES, MCP-1, CCR2, CCR5, CXCR1 and CXCR4 gene polymorphisms are not associated with the outcome of hepatitis B virus infection: Results from a large scale single ethnic population. J Korean Med Sci 22: 529-535, 2007.
38. Galan JJ, De Felici M, Buch B, Rivero MC, Segura A, Royo JL, Cruz N, Real LM and Ruiz A: Association of genetic markers within the KIT and KITLG genes with human male infertility. Hum Reprod 21: 3185-3192, 2006.
39. Dupont WD and Plummer WD Jr: Power and sample size calculations for studies involving linear regression. Control Clin Trials 19: 589-601, 1998.
40. Mehta SA, Christopherson KW, Bhat-Nakshatri P, Goulet RJ Jr, Broxmeyer HE, Kopelovich L and Nakshatri H: Negative regulation of chemokine receptor CXCR4 by tumor suppressor p53 in breast cancer cells: Implications of p53 mutation or isoform expression on breast cancer cell invasion. Oncogene 26: 3329-3337, 2007.
41. Jin B, Jiang F and Ding Z: Revaluation of the association between vascular endothelial growth factor gene 936 C/T polymorphism and breast cancer risk. Breast Cancer Res Treat 128: 909-912, 2011.
42. Igartua C, Myers RA, Mathias RA, Pino-Yanes M, Eng C, Graves PE, Levin AM, Del-Rio-Navarro BE, Jackson DJ, Livne OE, *et al*: Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. Nat Commun 6: 5965, 2015.
43. Su G, Christensen OF, Ostersen T, Henryon M and Lund MS: Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One 7: e45293, 2012.