Analysis of Disease Susceptibility Using Particle Swarm Optimization - Time Varying Acceleration to Generate SNP-barcode

Cheng-Hong Yang, *Member, IAENG*, Yu-Da Lin, and Li-Yeh Chuang

***Abstract*—Currently, an important and challenging task in genetic associations studies is the identification of common complex multi-factorials for diseases susceptibility. Given the significant computational association between SNPs with genotypes (SNP barcodes), current statistical methods have difficulty computing all possible combinations of SNPs with genotypes. This study proposes an improved particle swarm optimization (PSO), which is combined with the time varying acceleration method to overcome this challenge. The proposed method, called PSO-TVAC, is used to compute the association of genotype frequencies of case and control data based on statistical analysis. We systematically evaluated the method on the combined effect of 19 SNPs from seven published oxidative damage repair-related genes involved in breast cancer-related pathways. Odds ratio and risk ratio analyses are used to estimate SNP barcodes with significant differences between controls and cases. The estimated *OR* of the best SNP combination with genotypes (called the SNP barcode) is significantly greater than 1 (between 1.11 and 1.61) for specific combinations of two to seven SNPs in high risk groups. The results show that PSO-TVAC successfully improves on the inherent disadvantages of PSO for the identification of high-order SNP barcodes.**

***Index Terms*—SNP-SNP interaction, Particle Swarm Optimization, Time varying acceleration.**

I. Introduction

Genome-wide case-control association studies (GWAS) have been widely used to identify a set of single nucleotide polymorphisms (SNPs) to determine which are most closely associated with disease and cancer [1-4]. Many studies have hypothesized that the risk of disease and cancer is associated with the co-occurrence of SNPs on the genetic and phenotypic variability among individuals. The associations between cases and controls were found to significantly impact their susceptibility to disease and cancer. However, association studies for multiple SNP candidates remain computationally challenging. The "SNP barcode" used in this study can be regarded as an SNP genotype combination, e.g., TT, TC and CC for an SNP with a T/C polymorphism.

The possible combinations of SNP barcodes between cases and controls can be computed as C(X,Y)*3^Y, where X is the total number of SNPs, and Y is the number of selected SNPs. Many computational approaches have been proposed to examine epistasis in family-based and case-control association studies [5-10]. However, these methods were not sufficiently robust to simultaneously evaluate the complex interactions for all SNPs in several genes. Evolutionary algorithms (e.g., particle swarm optimization (PSO) [11] and genetic algorithms (GA) [12]) have been shown to be effective in reducing the number of search items among a greater number of SNP combinations. However, the PSO and GA methods do not guarantee that every implemented result contains a relevant solution when the SNP number is excessively large.

In this study, 24 SNPs obtained from seven oxidative damage repair-related genes (CAT, GPX1, GPX4, GSR, SOD2, TXN, and TXNRD2) which had been used to investigate single-factor association with breast cancer [13], were used to analyze multi-factor association with breast cancer. Our previous study [13] determined the effect of individual SNPs, but did not investigate their association with SNPs. However, analysis of association with SNPs might provide further insight into disease susceptibility.

We hypothesize that the interactions between polymorphisms of oxidative damage repair-related genes could have a synergistic effect on the pathogenesis of breast cancer, and differences between cases and controls can explain interactions in disease susceptibility. We propose the PSO-TVAC method to generate a SNP barcode to analyze the risk factors of disease susceptibility. The best combination of SNPs with genotypes can be verified by computing the odds ratio (*OR*) and its confidence intervals. We systematically evaluate the combination effects of 24 SNPs from seven oxidative damage repair-related genes involved in breast cancer. The SNP barcode generated by the PSO-TVAC algorithm is found to be statistically significant in predicting susceptibility to breast cancer, and the identified differences between cases and controls were an improvement over the PSO algorithm.

II. Method

A. Particle Swarm Optimization (PSO)

The particle swarm optimization algorithm (PSO) was

C.H. Yang is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, 80778, Kaohsiung Taiwan (phone: 886-7-3814526#5639; E-mail: chyang@cc.kuas.edu.tw).

Y.D. Lin is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, 80778, Kaohsiung, Taiwan (E-mail: e0955767257@yahoo.com.tw).

L.Y. Chuang is with the Department of Chemical Engineering, I-Shou University , 84001 , Kaohsiung, Taiwan (E-mail: chuang@isu.edu.tw).

developed by Kennedy and Eberhart [14] as an evolutionary computation algorithm that simulates social behavior based on information exchange. In PSO, the solution to the problem can be found in the particles of population. Each particle adjusts its vector according to its experience and the swarm experience to converge on a location to search for a good solution. The basic elements of PSO are described below:

1) Population: A population consists of *N* particles.

2) Particle vector, x_i: A solution can be represented by a *D*-dimensional vector; the *i*th particle can be described as $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, where x_{iD} is a *D*th dimensional value.

3) Particle velocity, v_i: The velocity of the *i*th particle is represented by $v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$, where v_{iD} is a velocity value in the *D*th dimension. In addition, each velocity must be limited within $[V_{\min}, V_{\max}]^D$.

4) Inertia weight, *w*: The *w* controls the impact of the particle's previous velocity on its current velocity. This control parameter affects the trade-off between the particle's exploration and exploitation abilities.

5) Individual best value, *pbest$_i$*: *pbest$_i$* is the vector of the *i*th particle with the highest fitness value at a given iteration.

6) Global best value, *gbest*: The *gbest* is the best vector amongst the particles' *pbest*.

7) Termination criteria: Stop condition of the PSO procedure.

The PSO procedure can be divided into four steps. First, the position and velocity of each particle in the population was randomly generated. Second, the *pbest$_i$* for each particle was updated by comparing its current fitness to the fitness of *pbest*. Third, the common knowledge (i.e., *gbest*) was updated according to the best *pbest* amongst the population. Fourth, the position of each particle was updated according its *pbest* and *gbest*. The update equations can be formulated as:

$$\begin{aligned} v_{id}^{new} &= w \times v_{id}^{old} + c_1 \times r_1 \times \left(pbest_{id} - x_{id}^{old}\right) \\ &\quad + c_2 \times r_2 \times \left(gbest_d - x_{id}^{old}\right) \end{aligned} \tag{1}$$

$$x_{id}^{new} = x_{id}^{old} + v_{id}^{new} \tag{2}$$

where *w* is the inertia weight which is a positive linear function of time that changes with the generations, r_1 and r_2 are random numbers between (0, 1), and c_1 and c_2 are acceleration constants that control how far a particle moves in a single generation. Velocities v_{id}^{new} and v_{id}^{old} respectively denote the new and old velocity of the particle, while x_{id}^{old} is the current particle position, and x_{id}^{new} is the updated particle position. The velocity implies the distance to which the particle's position should be moved in a generation, so that the velocity can move the particle towards the best position. The particles' velocities in each dimension were limited to within $[V_{\min}, V_{\max}]^D$, and the particles' positions are limited within $[X_{\min}, X_{\max}]^D$.

B. Particle swarm optimization - time varying acceleration coefficients (PSO-TVAC)

In PSO, the learning factors c_1 and c_2 represent the acceleration constants, which are usually equal to 2. c_1 and c_2 can influence the particle's search direction between the *pbest* and *gbest* vectors, in which c_1 controls exploitation search and c_2 controls exploration search. Unlike the PSO algorithm, the idea of PSO-TVAC is that c_1 and c_2 can be adjusted through the iteration number. The factor c_1 decreases from 2.5 to 0.5 through the iterations, while the factor c_2 increases from 0.5 to 2.5. The linear adjustments of c_1 and c_2 are defined in Eqs. 3 and 4:

$$c_1 = \left(c_{1\max} - c_{1\min}\right) \times \left(\frac{iteration}{iteration\_\max}\right) + c_{1\min} \tag{3}$$

$$c_2 = \left(c_{2\max} - c_{2\min}\right) \times \left(\frac{iteration}{iteration\_\max}\right) + c_{2\min} \tag{4}$$

where $c_{1\max}$ and $c_{1\min}$ respectively express the initial and maximal values in c_1, while $c_{2\max}$ and $c_{2\min}$ respectively express the initial and maximal values in c_2, *iteration* is the present iteration number, and *iteration*_max is the maximal iteration number. Therefore, the TVAC method can adjust the search behavior from exploitation to exploration, thus preventing particle dispersion and early convergence.

C. Application of the PSO-TVAC algorithm

a) Encoding schemes

In PSO-TVAC, the particle of population is defined as a vector divided into two parts: the selected SNPs and their genotypes, in which SNPs cannot be repeatedly selected. The particle encoding can be represented as follows:

$$Particle_i = (SNP_{i,j}, genotype_{i,j}), i=1, 2\ldots m, j=1, 2,\ldots, n$$

where $SNP_{i,j}$ represents the selected SNPs, $genotype_{i,j}$ represents the genotypes (i.e., types AA, Aa, and aa) once $SNP_{i,j}$ is selected, *m* is the size of the population and *n* is the number of SNPs selected. In the initialization step, the particle is randomly generated in each dimension. For example, the initial particle is randomly generated and represented as *particle* = ($SNP_{2,4,6}$, $Genotype_{1,2,3}$). In the particle, $SNP_{3,4,8}$ represents the chosen SNPs (2,4,6), and $Genotype_{1,2,3}$ represents the chosen genotypes (1,2,3). In this case, selected SNPs and SNPs associated with the genotypes were as follows: (2,1), (4,2) and (6,3).

b) Fitness function

A fitness value was used to compute the difference between cases and controls from the SNP barcode. The goal is to determine the highest fitness value, i.e., the maximum difference between cases and controls. The concept of fitness uses the intersection of set theory to compute the case and control sets that contain all elements of one of these sets that also belong to the other, but no other elements. The relevant equation is as follows:

$$\begin{aligned} &F(particle_i) = \\ &\quad \text{number}(C \cap particle_i) - \text{number}(N \cap particle_i) \end{aligned} \tag{5}$$

where the number( ) symbol denotes the total number of elements in a set, C denotes the total number of SNP barcodes in the case group, N represents the total number of SNP barcodes in the control group, and $particle_i$ represents the i^{th} particle. The fitness function can be divided into three separate steps. First, the total number of intersections of the cases and i^{th} particle is calculated as number($C \cap particle_i$). Second, the total number of intersections of the controls and i^{th} particle is calculated as number ($N \cap particle_i$). Finally, Eq. 5 is used to calculate the fitness value of the intersections of the cases and controls.

For example, we assume a particle= $(SNP_{2,5}, genotype_{1,3})$ to compute the fitness. First, we calculate the control number for SNP_2 with genotype 1 and SNP_5 with genotype 3. The number of cases matching the SNP_2 with genotype 1 includes 171 samples in the breast cancer data. Second, we calculate the number of controls independently matching the SNP_5 with genotype 3 as including 121 samples. According to Eq. (5), the fitness value is calculated by subtracting 171 from 121, giving 50.

c) Update *pbest* and *gbest*

Updating the *pbest* of particle and *gbest* of the population aims to move the particle toward a better search location. The particle attempts to find its best position (*pbest*) and the global best position (*gbest*). If the fitness value of a particle *P* in the current iteration is better than the fitness value of *pbest* in the previous iteration, then *pbest* is updated to *P*. If the fitness value of a particle's *pbest* in the current iteration is better than *gbest* in the previous iteration, then the *gbest* is updated to the *pbest*. The particle then adjusts its direction based on *pbest* and *gbest* in the following iteration.

The PSO-TVAC pseudo-code is shown below to describe how the algorithm collocates data through the above-mentioned procedure to obtain the best SNP barcode for breast cancer.

PSO-TVAC pseudo-code

```
Begin
Initialize population
While (number of iterations, or the stopping criterion is not met)
  Evaluate fitness of population via Eq. 5
  For n = 1 to number of particles
    Find pbest
    Find gbest
  For d = 1 to number of dimension of particle
     Update the position of particles via Eqs. 1 and 2
  Next d
  Next n
  Update the c_1 and c_2 value via Eqs. 3 and 4
Continue generation until stopping criterion is met
```

III. Result and Discussion

A. Parameter settings

Both termination conditions for both of PSO and PSO-TVAC are reached at the after 100 iterations. Population size is equal to 50. The sets of parameters c_1 and c_2 of PSO are equal to 2. In PSO-TVAC, both parameters c_{1_max} and c_{2_min} are equal to 2.5; both parameters c_{1_min} and c_{2_max} are equal to 0.5. V_{max} is equal to ($X_{max} - X_{min}$) and V_{min} is equal to – ($X_{max} - X_{min}$).

B. Data sets

The data sets consist of SNP genotype frequencies published in the literature [13]. The dataset was collected from the oxidative damage repair-related genes (65 SNPs for 11 genes) in the breast cancer association study. The genotype frequencies of our simulated data are identical to the original raw data for the genotypes frequencies [13]. Therefore, we used the SNP genotype frequencies to simulate the case and control data (5000 samples). The simulated data was randomly generated and thus obeyed the original genotype frequency in the whole dataset. We assumed that the amounts of three genotypes AA, Aa and aa in the SNP in the original data were 2132, 1970 and 449, respectively. We calculated the percentage of each genotype, i.e., 2132/4551 (47%) for AA, 1970/4551 (43%) for Aa and 449/4551 (10%) for aa. The simulated data for the SNP rs3020314 is then generated according to these three percentages, i.e., 47%× 5000=2350 for AA, 43%×5000=2150 for Aa and 10%×5000=500 for aa. The simulated data for the SNP has thus been controlled to 5000 (2350+2150+500=5000). The above procedure is shown in the "Pseudo-code for randomly generated data" below.

Pseudo-code for randomly generated data

```
begin
Set size = 5000
Set number of genotype = 3
Calculate amount of three genotypes
while (all SNPs are not controlled)
    Calculate amount of each genotype
    Calculate numbers of each controlled genotype
    for n = 1 to number of genotype
     Randomly create numbers of each controlled genotype
 next n
end
```

C. Performance measurement using statistical analysis

In this study, five statistical analysis were used to determine the SNP barcode [15] as below.

$$\text{Correct} = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

$$\text{Sensitivity} + \text{Specificity} = \frac{TP}{TP + FN} + \frac{TN}{FP + TN} \tag{7}$$

$$\text{Positive Predictive Value (PPV)} + \text{Negative Predictive Value (NPV)}$$
$$= \frac{TP}{TP + FP} + \frac{TN}{FN + TN} \tag{8}$$

$$\text{Risk Ratio } = \frac{TP \times (FP + TN)}{FP \times (TP + FN)} \tag{9}$$

$$\text{Odds Ratio} = \frac{TP \times TN}{FP \times FN} \tag{10}$$

TP is the number of cases matching the SNP barcode, *TN* is the number of controls not matching the SNP barcode, *FN* is the number of cases not matching the SNP barcode, and *FP* is the number of controls matching the SNP barcode. The risk ratio (*RR*) and odds ratio (*OR*) are used to measure the disease risk of the SNP barcode. *OR* is widely used in medical reports and offers a very convenient interpretation in case-control studies when the *RR* cannot be obtained directly (case-control association). The odds ratios are often interpreted as a *RR*. If the *OR* value is equal to one, the risk associated with the disease for the given SNP barcode is the same as the overall risk estimated from all cases and controls. A larger *OR* value (>1) indicates a risk association between the SNP barcode and the disease. Similarly, a lower *OR* value (<1) indicates a protective association between the SNP barcode and the disease.

D. Identification of best SNP barcode

As shown in Table I, among the order SNP barcodes three specific combined SNPs with genotypes (i.e., SNPs (4,12,20) with genotype 2-1-1 ; [rs757229-Aa]-[rs4135179-AA]-[rs2073752-AA]) showed the maximal difference between breast cancer and non-cancer groups. Similarly, two to seven best-performing combined-SNP barcodes are mined by PSO-TVAC (Table I) through the complete result set. The PSO-TVAC provides a good difference between the breast cancer and non-cancer groups with a fixed number of SNPs.

E. Analyzing the ranks of OR and RR for breast cancer

The *OR* was widely applied in medical reports and offers a very important interpretation in case-control studies. An *OR* value bigger than 1 indicates a stronger association between cases and controls for the risk of disease. Table I shows the risk association with specific SNP barcodes and other combinations (two to seven) in breast cancer. The *OR* values (1.11-1.61) and the *RR* values (1.05-1.24) increase with the number of SNP combinations in high risk cases. We observed that the PSO-TVAC (Table I) provides higher *OR* values (1.11-1.61) with two to seven order SNP barcodes for the risk of breast cancer. The number of case groups was greater than the number of control groups, which means the SNP barcode can influence the risk of breast cancer. Thus, women with the specific SNP barcode implied a risk ratio that represented significantly increased *OR* values of 1.11-1.61 for breast cancer. The results suggested that genes with these SNP barcodes represent a risk for breast cancer. On the other hand, Table II showed the PSO only identified an available SNP barcode in 2-SNP which was determined by the *p*-value. The *p*-value represented the confidence level of results analysis and, to be significant, a result has to be over 0.05. PSO (Table II) only provides the *OR* (1.11-1.22) values with two to seven order SNP barcodes for the risk of breast cancer. Therefore, the PSO-TVAC provided a better differentiation in terms of association of the fixed SNP barcode between the cases and controls.

F. Comparing PSO-TVAC with PSO for SNP-SNP interaction in breast cancer

In this study, we investigated the association for case-control studies with multiple-SNPs to analyze 24 SNPs obtained from seven oxidative damage repair-related genes in breast cancer. The SNPs involved in the analysis of association studies were difficult to compute, especially the very high-order SNPs which were also investigated. We proposed a PSO-TVAC algorithm to perform a powerful identification of SNP-SNP interactions for breast cancer. The statistical methods, such as *p*-value, *OR* and its 95% CI, provided strong evidence to explain the ability of PSO-TVAC to identify the best difference between cases and controls.

Tables I and II show the combinations of 2- to 7-SNPs with their associated genotypes. The results were compared with the differences between cases and controls. The combination of 2-SNP with their corresponding genotypes, SNPs (4, 12) with genotype 2-1, [rs757229-Aa]-[rs4135179-AA], were identified as having 107 differences between the case group and control group (1581 *vs.* 1474) by PSO-TVAC and PSO. The results for 3- to 7-SNPs clearly show that the PSO-TVAC exhibited superior searching ability to that of the PSO in terms of comparison between the cases and controls. For example, in a 3-SNP combination, the combination consists of SNPs (4, 12, 20) with genotype 2-1-1 ([rs757229-Aa]-[rs4135179-AA]-[rs2073752-AA]), which was identified as having an 89 difference value by PSO-TVAC. On the other hand, PSO identified the combination consisting of SNPs (12, 13, 19) with genotype 1-2-2 ([rs4135179- AA]-[rs2301241-Aa]-[rs1548357-Aa]) as having a difference value of 49.

The time varying acceleration can be observed in Eq.1; it was only used to change the values of c_1 and c_2 in the original PSO updating equation. The computational complexities of PSO and PSO-TVAC can be estimated by the fitness function computation. We defined *i* and *p* respectively as the number of iterations and the number of particles. The fitness function computation can then be represented as the computational complexity of O(*ip*), in which O( ) was the big O notation.

IV. Conclusion

In individuals, the genetic genes can influence the risk of developing many diseases or cancers. Therefore, the effect of multiple factor association on disease susceptibility is widely used in genome-wide case-control association studies. Large-scale SNPs analysis of association studies are difficult to perform, especially when multiple SNPs are investigated simultaneously. This study proposed the PSO-TVAC to identify 24 SNP cross-interactions and provide representative gene-gene interactions for breast cancer. The time varying acceleration method was successfully used to improve PSO to provide significant searches within limited time frames, thus enhancing the opportunity to obtain maximum difference between cases and controls in higher-order SNP-SNP interactions. Results involving two- to seven-SNPs show the *OR* of the best SNP barcodes is in the range of 1.11 to 1.61, and the 95% CI of the *OR* is in the range of 0.98 to 2.66. All SNP barcodes show significantly reduced *OR* values (*p*-value < 0.050 to 0.001). This suggests that the PSO-TVAC method is suitable for the systematic exploration of genome-wide SNP interactions.

Table I
ESTIMATED BEST SNP COMBINATIONS ON THE OCCURRENCE OF BREAST CANCER BY TVAC-PSO

Combined SNP number (specific SNPs)	SNP Genotypes	Control number / Case number	Correct	Sen. + Spe.	PPV+NPV	Risk Ratio	Odds Ratio (95%CI)	*p-value*
2-SNP	others	3419/3526	0.51	1.02	1.02	1.05	1.11	0.02
SNPs(4-12)	2-1	1581/1474				(1.01-1.10)	(1.02-1.21)	
3-SNP	others	4149/4238	0.51	1.02	1.03	0.07	1.14	0.02
SPNs(4-12-20)	2-1-1	851/762				(1.01-1.12)	(1.02-1.27)	
4-SNP	Others	4624/4684	0.51	1.01	1.04	1.09	1.21	0.02
SPNs(4-12-14-20)	2-1-1-1	376/316				(1.01-1.17)	(1.03-1.41)	
5-SNP	Others	4812/4847	0.50	1.01	1.05	1.11	1.24	0.05
SPNs(4-8-12-14-20)	2-2-1-1-1	188/153				(1.00-1.22)	(1.00-1.55)	
6-SNP	Others	4912/4937	0.50	1.00	1.08	1.17	1.40	0.04
SPNs(5-8-9-14-20-23)	2-2-2-1-1-2	88/63				(1.00-1.33)	(1.00-1.97)	
7-SNP	Others	4955/4972	0.50	1.00	1.12	1.24	1.61	0.05
SPNs(5-8-9-10-14-20-23)	2-2-2-1-1-1-2	45/28				(0.99-1.46)	(0.98-2.66)	

*The SNP combinations on the occurrence of breast cancer have significant (*p*-value < 0.05).

Table II
ESTIMATED BEST SNP COMBINATIONS ON THE OCCURRENCE OF BREAST CANCER BY PSO

Combined SNP number (specific SNPs)	SNP Genotypes	Case number / Control number	Correct	Sen. + Spe.	PPV+NPV	Risk Ratio	Odds Ratio (95%CI)	*p-value*
2-SNP	others	3419/3526	0.51	1.02	1.02	1.05	1.11	0.02
SNPs(4-12)	2-1	1581/1474				(1.01-1.10)	(1.02-1.21)	
3-SNP	others	4340/4389	0.51	1.01	1.02	1.04	1.09	0.14
SPNs(12-13-19)	1-2-2	660/611				(0.98-1.11)	(0.97-1.23)	
4-SNP	Others	4583/4610	0.50	1.00	1.01	1.04	1.08	0.32
SPNs(6-12-17-20)	2-1-2-1	417/390				(0.96-1.11)	(0.93-1.25)	
5-SNP	Others	4869/4892	0.50	1.00	1.04	1.10	1.22	0.13
SPNs(4-5-8-14-16)	2-2-2-1-1	131/108				(0.97-1.23)	(0.93-1.59)	
6-SNP	Others	4924/4933	0.50	0.99	1.03	1.06	1.14	0.45
SPNs(3-6-8-10-14-16)	1-2-2-1-1-2	76/67				(0.89-1.23)	(0.81-1.60)	
7-SNP	Others	4955/4959	0.50	0.99	1.02	1.05	1.10	0.67
SPNs(4-10-15-17-20-22-23)	2-1-2-2-1-1-2	45/41				(0.83-1.26)	(0.70-1.72)	

*The SNP combinations on the occurrence of breast cancer have significant (p-value < 0.05).

Acknowledgment

This work is partly supported by the National Science Council in Taiwan under grants 101-2622-E-151-027-CC3, 100-2221-E-151-049-MY3, and 100-2221-E-151-051-MY2.

REFERENCES

[1] X. Li, H. Chen, J. Li, and Z. Zhang, "Gene function prediction with gene interaction networks: a context graph kernel approach," *IEEE Transactions on Information Technology in Biomedicine,* vol. 14, pp. 119-128, 2010.

[2] P. Kraft and C. A. Haiman, "GWAS identifies a common breast cancer risk allele among BRCA1 carriers," *Nat Genet,* vol. 42, pp. 819-20, 2010.

[3] D. Fanale, V. Amodeo, L. R. Corsini, S. Rizzo, V. Bazan, and A. Russo, "Breast cancer genome-wide association studies: there is strength in numbers," *Oncogene,* vol. 31, pp. 2121–2128, 2011.

[4] J. C. Yu, C. N. Hsiung, H. M. Hsu, B. Y. Bao, S. T. Chen, G. C. Hsu, W. C. Chou, L. Y. Hu, S. L. Ding, C. W. Cheng, P. E. Wu, and C. Y. Shen, "Genetic variation in the genome-wide predicted estrogen response element-related sequences is associated with breast cancer development," *Breast Cancer Res,* vol. 13, pp. R13, 2011.

[5] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics,* vol. 26, pp. 445-455, 2010.

[6] C. H. Yang, L. Y. Chuang, Y. J. Chen, H. F. Tseng, and H. W. Chang, "Computational analysis of simulated SNP interactions between 26 growth factor-related genes in a breast cancer association study," *OMICS,* vol. 15, pp. 399-407, 2011.

[7] P. Yang, J. W. Ho, Y. H. Yang, and B. B. Zhou, "Gene-gene interaction filtering with ensemble of filters," *BMC Bioinformatics,* vol. 12, pp. S10, 2011.

[8] 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, pp. e37018, 2012.

[9] L. Y. Chuang, H. W. Chang, M. C. Lin, and C. H. Yang, "Chaotic particle swarm optimization for detecting SNP-SNP interactions for CXCL12-related genes in breast cancer prevention," *European Journal of Cancer Prevention,* vol. 21, pp. 336-342, 2012.

[10] 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 Journal of Medical Sciences,* vol. 28, pp. 362-368, 2012.

[11] C. H. Yang, H. W. Chang, Y. H. Cheng, and L. Y. Chuang, "Novel generating protective single nucleotide polymorphism barcode for breast cancer using particle swarm optimization," *Cancer epidemiology,* vol. 33, pp. 147-154, 2009.

[12] C. H. Yang, L. Y. Chuang, Y. J. Chen, H. F. Tseng, and H. W. Chang, "Computational Analysis of Simulated SNP Interactions Between 26 Growth Factor-Related Genes in a Breast Cancer Association Study," *OMICS: A Journal of Integrative Biology,* vol. 15, pp. 399-407, 2011.

[13] P. D. P. Pharoah, J. Tyrer, A. M. Dunning, D. F. Easton, B. A. J. Ponder, and S. Investigators, "Association between common variation in 120 candidate genes and breast cancer risk," *PLoS Genetics,* vol. 3, pp. 401-406, 2007.

[14] J. Kennedy and R. Eberhart, "Particle swarm optimization," *IEEE International Joint Conference on Neural Network*, pp. 1942-1948, 1995.

[15] L. E. Mechanic, B. T. Luke, J. E. Goodman, S. J. Chanock, and C. C. Harris, "Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions," *BMC Bioinformatics,* vol. 9, pp. 146, 2008.