

Original Paper

Identification of High-Order Single-Nucleotide Polymorphism Barcodes in Breast Cancer Using a Hybrid Taguchi-Genetic Algorithm: Case-Control Study

Li-Yeh Chuang¹, PhD; Cheng-San Yang², MD, PhD; Huai-Shuo Yang³, MA; Cheng-Hong Yang^{3,4,5}, PhD

¹I-Shou University, Kaohsiung City, Taiwan

²Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi City, Taiwan

³Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung City, Taiwan

⁴Drug Development and Value Creation Research Center, Kaohsiung Medical University, Kaohsiung, Taiwan

⁵College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

Corresponding Author:

Cheng-Hong Yang, PhD

Department of Electronic Engineering

National Kaohsiung University of Science and Technology

No. 415 Jiangong Road, San-Min District

Kaohsiung City, 82778

Taiwan

Phone: 886 7 381 4526

Email: chyang@nkust.edu.tw

Abstract

Background: Breast cancer has a major disease burden in the female population, and it is a highly genome-associated human disease. However, in genetic studies of complex diseases, modern geneticists face challenges in detecting interactions among loci.

Objective: This study aimed to investigate whether variations of single-nucleotide polymorphisms (SNPs) are associated with histopathological tumor characteristics in breast cancer patients.

Methods: A hybrid Taguchi-genetic algorithm (HTGA) was proposed to identify the high-order SNP barcodes in a breast cancer case-control study. A Taguchi method was used to enhance a genetic algorithm (GA) for identifying high-order SNP barcodes. The Taguchi method was integrated into the GA after the crossover operations in order to optimize the generated offspring systematically for enhancing the GA search ability.

Results: The proposed HTGA effectively converged to a promising region within the problem space and provided excellent SNP barcode identification. Regression analysis was used to validate the association between breast cancer and the identified high-order SNP barcodes. The maximum OR was less than 1 (range 0.870-0.755) for two- to seven-order SNP barcodes.

Conclusions: We systematically evaluated the interaction effects of 26 SNPs within growth factor-related genes for breast carcinogenesis pathways. The HTGA could successfully identify relevant high-order SNP barcodes by evaluating the differences between cases and controls. The validation results showed that the HTGA can provide better fitness values as compared with other methods for the identification of high-order SNP barcodes using breast cancer case-control data sets.

(*JMIR Med Inform* 2020;8(6):e16886) doi: [10.2196/16886](https://doi.org/10.2196/16886)

KEYWORDS

genetic algorithm; single-nucleotide polymorphism; breast cancer; case-control study

Introduction

Breast cancer has a major disease burden in the female population, with a growing incidence recently [1,2]. Previously, several interpretations of associations between breast cancer

and tumor characteristics [3-5], single-nucleotide polymorphisms (SNPs) [6-8], clinicopathological factors [9], and biomarkers [10] revealed relevant association effects between these factors and the risk of cancer. Previous studies also indicated that genomic variation could contribute to the

tumorigenicity process in breast cancer [11-14]. Thus, effective approaches for breast cancer estimation are required.

SNPs are crucial genetic variants in genomic association analyses involving leukemia [15], cancers [16], and other diseases [17-19]. Numerous SNPs cannot be excluded from analyses as no relevant differences between cases and controls can be found through conventional methods. Some SNPs may have relevant associations with other SNPs, and these associations are referred to as SNP barcodes. Consequently, the detection of SNP barcodes is vital for association analyses of diseases and cancers [20-23].

An SNP barcode consists of SNPs, and each SNP includes three genotypes. The large space of suitable SNP barcode combinations complicates the statistical evaluation and identification of relevant SNP barcodes. Evolutionary algorithms have been proposed to facilitate statistical identification of SNP barcodes, and a genetic algorithm (GA) is one of the most frequently used algorithms in genomic studies [24,25]. A GA is an effective approach in the identification of relevant genetic associations for various diseases through the use of more efficient search abilities to enhance population diversity [26]. The crossover and local search operations in a GA can reduce the probability of the same vector being identified between two selected SNPs, and hence, they can improve the search ability of this algorithm.

Breast cancer is a major health issue, and machine learning algorithms are frequently employed to detect the complex genomic associations in breast cancer studies. Although previous machine learning approaches could effectively identify SNP associations in genomic studies, the detection rate of SNP barcodes remains challenging for high-order SNP barcodes. Thus, we proposed a hybrid Taguchi-genetic algorithm (HTGA) for high-order SNP barcode identification in a breast cancer case-control study.

Methods

Genetic Algorithm

A GA is a machine learning algorithm inspired by biological evolutionary processes [27]. The first GA operation is population initialization, in which solutions are produced over the solution space; these initial solutions are designated as parents. In the population, two parents are strategically selected according to some fitness values for crossover operators. Crossover operators generate offspring by combining the chromosomal matter from

the two parents. Mutation operations can increase population diversity through localized change, eliminating inferior chromosomes from the population and retaining good offspring. Thus, the good factors within the population can be passed on to the next generation. The aforementioned operations and population replacement are repeated until the stopping criterion is satisfied.

Taguchi Method

The methods proposed by Taguchi et al [28] are based on a statistical experimental design to improve the evaluation and performance of products, process conditions, and parameter settings. Taguchi methods primarily rely on orthogonal arrays (OAs) and the signal-to-noise ratio (SNR). An OA is a fractional factorial matrix that provides a comprehensive analysis of interactions among all design factors. This matrix ensures a proportionate comparison of levels for all factors. A two-level OA can be defined as $L_n(2^{n-1})$, where $n=2^k$ is the number of experimental runs, k (1) is a positive integer, base 2 represents two levels for each design parameter, and $n-1$ is the number of columns in the OA. "L" represents "Latin," because the OA experimental design concept is associated with the Latin square. An example of an OA is shown in Table 1.

SNR (η) is used as the selection quality characteristic in the field of communications engineering; it can be used to optimize the parameters for a target. Taguchi methods can classify the parameter design problem into several categories according to the problem. Both smaller-the-better and larger-the-better SNR types are used. Considering the set of characteristics y_1, y_2, \dots, y_n , in the smaller-the-better case, the SNR can be determined using the following equation:

$$\eta = -10 \log \left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) \quad (1)$$

In the larger-the-better case, the SNR can be determined using the following equation:

$$\eta = -10 \log \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{y_i^2} \right) \quad (2)$$

The SNR evaluates the robustness of the levels of each design parameter. A high-quality result can be achieved for a particular target by controlling the parameters at a particular level with a high SNR value.

Table 1. A L8(27) orthogonal array.

Experiment number	Factors						
	A	B	C	D	E	F	G
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

Hybrid Taguchi-Genetic Algorithm

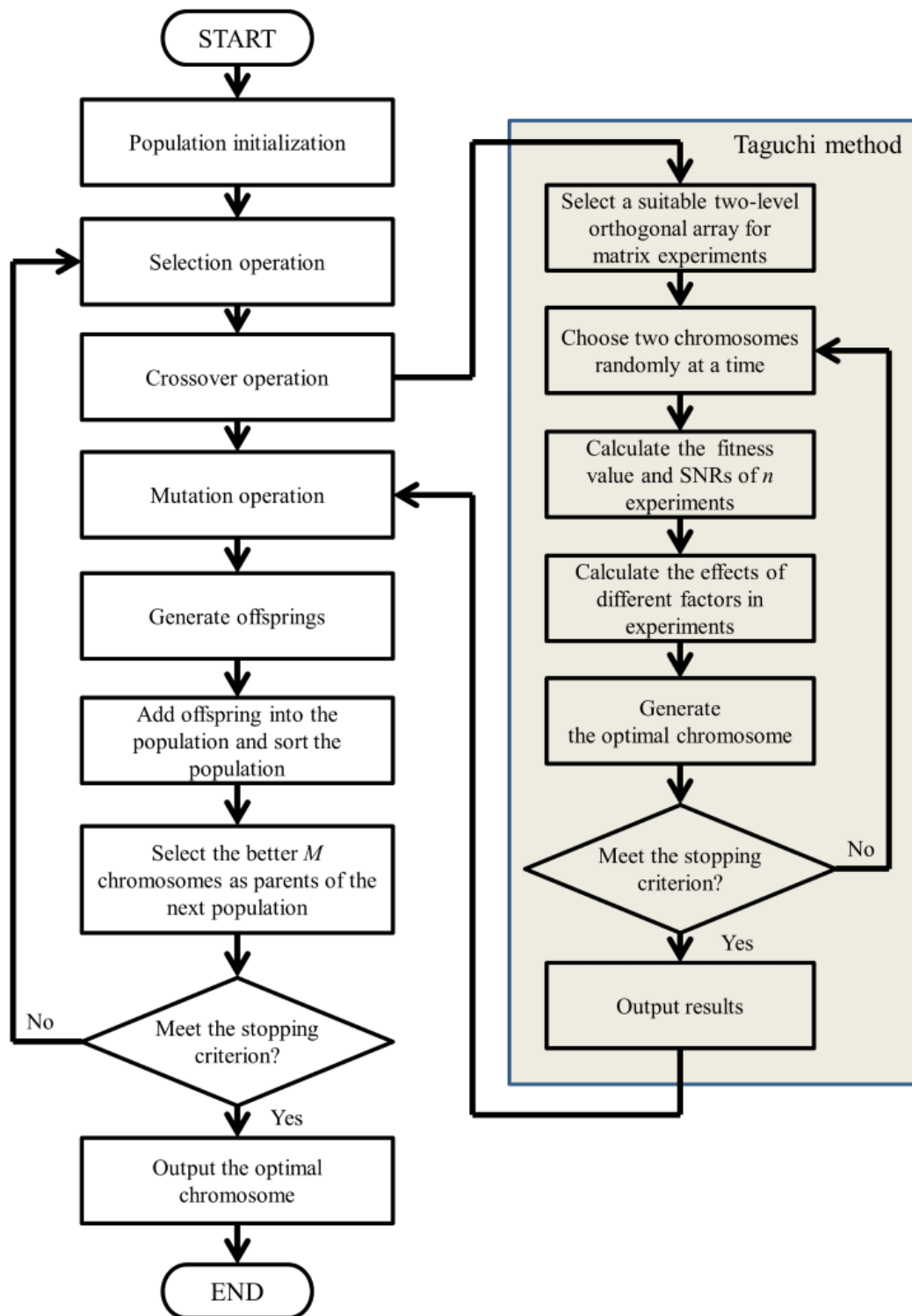
In the HTGA, a Taguchi method is added into GA crossover and mutation operations. [Figure 1](#) depicts a flowchart of the HTGA approach, which includes the below-mentioned 17 steps. The pseudocode of the HTGA is shown in [Textbox 1](#).

HTGA Procedure

The procedure involves the following 17 steps: (1) Population initialization, execute the algorithm and generate an initial population; (2) Fitness value evaluations, evaluate the population's fitness values; (3) Selection operation, select candidates using the tournament approach; (4) Crossover operation, the probability of crossover is determined by the crossover rate p_c ; (5) Select a suitable two-level orthogonal

array for the experiment; (6) Randomly choose two chromosomes at a time to execute a matrix experiment; (7) Calculate the function values and SNRs of n experiments in the orthogonal array $L_n(2^{n-1})$; (8) Calculate the effects of different factors and in the experiment; (9) An optimal chromosome is generated; (10) Repeat steps 5 through 8 until the expected number $(1/2) \times M \times p_c$ is reached; (11) New chromosomes are generated through the Taguchi method; (12) Mutation operation, mutation probability is determined by the mutation rate p_m ; (13) Add chromosomes from a pool into the population; (14) Sort the population by fitness; (15) Select the fittest chromosomes as the new population for the next generation; (16) If the stopping criterion is met, execute step 17; if not, go back to step 2; (17) The chromosome with the highest fitness value is the HTGA solution.

Figure 1. Hybrid Taguchi-genetic algorithm flowchart. SNR: signal-to-noise ratio.



Textbox 1. Pseudocode of the hybrid Taguchi-genetic algorithm.

```

Input: maximum iteration as  $T$  (termination criterion)

population size  $M$ 
crossover rate  $P_c$ 
mutation rate  $P_m$ 

Output: optimal chromosome (the optimal solution)

begin
# Initialization
/* Initialize  $M$  chromosomes as population */
for iteration  $\leftarrow 1$  to  $T$  do
  # Selection operation
  for  $i \leftarrow 1$  to  $M$  do
    /* Randomly select two chromosomes from population as chromosome1 and
    chromosome2 */
    if  $fitness_1 \geq fitness_2$  do
      winner  $\leftarrow$  chromosome1
    end if
    else do
      winner  $\leftarrow$  chromosome2
    end if
    /* put winner into mating pool */
  end for
  # Crossover operation
  for  $i \leftarrow 1$  to  $(M / 2)$  do
    /* Sequentially select two chromosomes from mating pool as chromosome1
    and chromosome2 */
    if random()  $< P_c$  do
      crossover(chromosome1, chromosome2)
      /* generate two offspring */
    end if
    /* put two offspring into offspring pool */
  end for
  # Taguchi operation
  for  $i \leftarrow 1$  to  $(0.5 \times M \times P_c)$  do
    /* Randomly select two chromosomes from offspring pool as chromosome1
    and chromosome2 */
    Taguchi(chromosome1, chromosome2)
    /* generate one offspring */
    /* put one offspring into offspring pool */
  end for
  # Mutation operation
  for  $i \leftarrow 1$  to size of offspring pool do

```

```

for  $j \leftarrow 1$  to dimensions of chromosome do
  if random() <  $P_m$  do
    mutation(chromosome[j])
  end if
end for
/* generate one offspring */
/* put one offspring into offspring pool */
end for
# Replacement operation
/* Reserve best  $M$  chromosomes as new population from population and offspring
pool */
end for
/* Obtain optimal chromosome */
End

```

Encoding Schemes and Population Initialization

In the proposed GA, a suitable solution to a problem is denoted as chromosome $C = \{c_1, c_2, \dots, c_n\}$, and the encoding scheme aims to design suitable elements in a chromosome. In the SNP barcode problem, the elements in a chromosome include (1) the indexes of the selected SNPs in the data set and (2) the genotypes of these selected SNPs. Thus, a chromosome C_i is expressed as shown in equation 3.

$$C_i = (SNP_{i,s}, Genotype_{i,g}) \quad (3)$$

where $i = 1, 2, \dots, m$, and is the population size. $SNP_{i,s}$, where $s = 1, 2, \dots, n/2$, is a selected SNP dimension in which all SNPs are unrepeatable, and n is the SNP barcode order. $Genotype_{i,g}$ represents the three possible genotypes of the selected $SNP_{i,s}$, where $g = n/2 + 1, n/2 + 2, \dots, n$ is the selected genotype dimension. In the population initialization, all chromosomes are stochastically generated according to the encoding schemes.

Fitness Function Evaluation

The aim of SNP barcode identification is to detect relevant differences between cases and controls. To optimize the protective effect of the SNP combination, a fitness function is required for comparing cases and controls. A high difference between cases and controls indicates a high probability of detecting relevant SNP barcodes. In the proposed GA, a

chromosome is measured by the fitness function shown in equation 4.

$$F(C_i) = \text{number}(control \cap C_i) - \text{number}(case \cap C_i) \quad (4)$$

where *number* is the total number of elements in a set, *control* denotes the controls, *case* denotes the cases, and C_i is the i th chromosome. Thus, the number of intersections between the i th chromosome and the controls is calculated by *number* ($control \cap C_i$), and the number of intersections between the i th chromosome and the cases is calculated by *number* ($case \cap C_i$). Thereafter, we calculate the difference between *number* ($control \cap C_i$) and *number* ($case \cap C_i$) as the fitness value at C_i .

Selection Operation

In the selection operation, a random tournament selection scheme is used to pick each pair of parents from the population [29]. In tournament selection, two chromosomes are randomly selected to compare their individual fitness values. The chromosomes with better fitness values are inserted into the mating pool. According to the mechanism of tournament selection, the probability that the average fitness value of solutions in the mating pool is better than the average fitness value of the parent population is high. Chromosomes in the mating pool are selected for the crossover operation and used to produce offspring. [Textbox 2](#) provides the pseudocode of tournament selection. The selection operation is repeatedly executed until the maximum mating pool size is achieved.

Textbox 2. Tournament selection procedure.

```

Input: population, the list of chromosomes to select from
Input: chromosome1, the first randomly selected chromosome from population
Input: chromosome2, the second randomly selected chromosome from population
Input: fitness1, the fitness value of the first chromosome
Input: fitness2, the fitness value of the second chromosome
Output: winner: the chromosome with better fitness value in tournament
Output: mating pool: reserve the list of chromosomes to execute crossover operation
# Tournament selection
begin
for i ← 1 to size of mating pool do
  Randomly select two chromosomes from population
  if fitness1 ≥ fitness2 do
    winner ← chromosome1
  end if
  else do
    winner ← chromosome2
  end else
  put winner into mating pool
end for

```

Crossover Operation

After the selection operation, the crossover operation is implemented to create high-performing individuals. Two chromosomes are sequentially selected from the mating pool as a pair of parents, and then, the crossover operation is executed on them. The crossover operation uses a uniform crossover. Each bit in a chromosome is randomly generated as 0 or 1, and for 1, points are swapped between parent organisms; otherwise, points are not swapped. The encoding schemes establish a single point as an SNP locus with a corresponding genotype locus at the $j \cdot 2 + 1$ position, where $j = 1, 2, \dots, n/2$ is the index in the chromosome and n is the SNP barcode order. Therefore, $n/2$ bits are randomly generated, and both the $j \cdot 2 + 1$ genotype locus and j th bit representing an SNP are swapped in the parent organisms.

Taguchi Operation

An orthogonal array exhibits Q design factors. Each factor has two levels. An orthogonal array $L_n(2^{n-1})$ exhibits $n-1$ columns and n individual experiments corresponding to n rows, where $n = 2^k$ and $Q \leq n-1$; k is a positive integer, defined as an integer > 1 , and it is used for adjusting the number of experimental runs.

The SNR (η) is the mean square deviation of the fitness function. Let two values of η be $\eta_i = (y_i)^2$ and $\eta_i = -(y_i)^2$ (where is negative) in the case of a fitness function that is maximized (larger-the-better). Let y_i be the function evaluation value of experiment $i = 1, 2, 3, \dots, n$, where n is the number of experiments. The effect of factor f is defined as follows:

$$E_{fl} = \text{sum of } \eta_i \text{ for factor } f \text{ at level } l \text{ (5)}$$

where i is the experiment number, f is the factor name, and l is the level number.

Mutation Operation

The mutation operation aims to prevent the population from falling into local optima. In all suitable solutions, each offspring element has a chance to undergo a mutation operation. Each mutation position with a probability of mutation p_m generates a random number in $(0, 1)$. If the number is less than p_m at the i th element in an offspring specimen, the i th element will be mutated by a randomly generated possible value.

Replacement Operation

The replacement operation uses an individual to replace the weakest individual in the population. After the completion of the aforementioned operations, the offspring are added to the population, and then, all the parents and offspring are ranked based on their fitness values. Subsequently, top p chromosomes in the population size are selected as the new population for the next generation, where p is the population size.

Termination Condition

The HTGA operation is repeated in successive iterations until the stopping criterion is met. In this study, a maximum number of iterations was used to terminate HTGA operations.

Parameter Setting

This study compared the search effectiveness of the HTGA with that of standard GA, particle swarm optimization (PSO) [30],

and chaotic PSO (CPSO) [31] methods. PSO is a swarm intelligence algorithm that simulates the social behavior of organisms. In PSO, each individual represents a particle and considers a potential solution in the swarm population. In CPSO, chaotic theory is incorporated into PSO to increase the search space and enhance PSO performance. PSO and CPSO parameters include population size, iteration size, minimum and maximum inertial weights, and learning factors. In each method, the number of iterations was set to 1000, and the population size was 50 for the test data set. In PSO and CPSO, the minimum and maximum inertial weights were 0.4 and 0.9, respectively. Both weights of learning factors c_1 and c_2 were set to 2. In the tested GA and the proposed HTGA, the probability of crossover (p_c) with an exchange probability was 0.3 and the probability of mutation (p_m) with an exchange probability was 0.05.

Statistical Analysis

The OR was used to evaluate the risk of an SNP barcode [32], and it was defined as follows:

$$OR = (TP \times TN) / (FP \times FN) \quad (6)$$

where TP represents the number of true positives, TN represents the number of true negatives, FN represents the number of false negatives, and FP represents the number of false positives.

Results

Data Sets

A set of 26 SNPs related to growth factor genes was selected to simulate a data set. Several growth factor-related breast

cancer genes (*EGF*, *IGF1*, *IGF1R*, *IGF2*, *IGFBP3*, *IL10*, *TGFBI*, and *VEGF*), including 26 SNPs, were used as simulation data to evaluate existing algorithms and the proposed HTGA. The data set only provided the genotype frequencies of each SNP without the original raw data of genotypes. Table 2 presents the SNPs and genotype distributions. The simulated frequencies of SNPs were acquired from the literature [33]. SNPs used in the original data comprised different numbers of individuals; therefore, the number of every SNP must be normalized to the same number. The new data were randomly generated according to the frequency of the original data. All SNP data from the data source were adjusted to 5000 samples for all genotype distributions. For example, for *SNP1* (gene, *EGF*; dBSNP ref. rs2237054), the total number of three genotypes (ie, TT, TA, and AA) in the control was 2273 (2008 + 259 + 6). The percentage for each genotype in *SNP1* was calculated as “original data*/sum (%)” (ie, 2008/2273, 88.3% for TT; 259/2273, 11.4% for TA; and 6/2273, 0.3% for AA), where the symbol “*” indicates that the original data were derived from the SNP data set before normalization. On the basis of this percentage, the modified data for *SNP1* were calculated by multiplying the percentage with the sum of the complete data set (SNP number adjusted to 5000) (ie, 88.3% × 5000 [n=4418] for AA; 11.4% × 5000 [n=569] for Aa, and 0.3% × 5000 [n=13] for aa). Therefore, the modified data for *SNP1* were adjusted to a total of 5000 (4418 + 569 + 13 = 5000). Thus, 5000 simulation samples of SNP genotypes were randomly generated by following fixed distribution.

Table 2. Estimated effect from individual single-nucleotide polymorphisms of 26 growth factor-related genes for the occurrence of breast cancer.

SNP ^a (gene)	SNP type	Case number/normal number	OR	95% CI	P value
1. rs2237054 (<i>EGF</i>)	1-TT	4408/4418	N/A ^b	N/A	N/A
1. rs2237054 (<i>EGF</i>)	2-TA	570/569	1	0.89-1.14	.97
1. rs2237054 (<i>EGF</i>)	3-AA	22/13	1.7	0.85-3.37	.18
2. rs5742678 (<i>IGF1</i>)	1-CC	2797/2866	N/A	N/A	N/A
2. rs5742678 (<i>IGF1</i>)	2-CG	1844/1837	1.03	0.95-1.12	.52
2. rs5742678 (<i>IGF1</i>)	3-GG	359/297	1.24	1.05-1.46	.01
3. rs1549593 (<i>IGF1</i>)	1-CC	2924/2970	N/A	N/A	N/A
3. rs1549593 (<i>IGF1</i>)	2-CA	1753/1771	1.01	0.93-1.09	.92
3. rs1549593 (<i>IGF1</i>)	3-AA	323/259	1.27	1.07-1.50	.008
4. rs6220 (<i>IGF1</i>)	1-AA	2643/2698	N/A	N/A	N/A
4. rs6220 (<i>IGF1</i>)	2-AG	1933/1951	1.01	0.93-1.10	.80
4. rs6220 (<i>IGF1</i>)	3-GG	424/351	1.23	1.06-1.44	.007
5. rs2946834 (<i>IGF1</i>)	1-CC	2295/2336	N/A	N/A	N/A
5. rs2946834 (<i>IGF1</i>)	2-CT	2171/2150	1.03	0.95-1.12	.53
5. rs2946834 (<i>IGF1</i>)	3-TT	534/514	1.06	0.93-1.21	.43
6. rs1568502 (<i>IGF1R</i>)	1-AA	2914/2955	N/A	N/A	N/A
6. rs1568502 (<i>IGF1R</i>)	2-AG	1840/1807	1.03	0.95-1.12	.46
6. rs1568502 (<i>IGF1R</i>)	3-GG	246/238	1.05	0.87-1.26	.65
7. IGF1R-10 (<i>IGF1R</i>)	1-AA	3169/3201	N/A	N/A	N/A
7. IGF1R-10 (<i>IGF1R</i>)	2-Aa	1545/1582	0.99	0.91-1.08	.77
7. IGF1R-10 (<i>IGF1R</i>)	3-aa	286/217	1.33	1.11-1.60	.003
8. rs2229765 (<i>IGF1R</i>)	1-GG	1523/1429	N/A	N/A	N/A
8. rs2229765 (<i>IGF1R</i>)	2-GA	2533/2489	0.96	0.87-1.05	.33
8. rs2229765 (<i>IGF1R</i>)	3-AA	944/1082	0.82	0.73-0.92 ^c	.001
9. rs8030950 (<i>IGF1R</i>)	1-CC	2737/2745	N/A	N/A	N/A
9. rs8030950 (<i>IGF1R</i>)	2-CA	1902/1917	1	0.92-1.08	.92
9. rs8030950 (<i>IGF1R</i>)	3-AA	361/338	1.07	0.92-1.25	.41
10. rs680 (<i>IGF2</i>)	1-GG	2538/2451	N/A	N/A	N/A
10. rs680 (<i>IGF2</i>)	2-GA	2074/2183	0.92	0.85-1.00	.04
10. rs680 (<i>IGF2</i>)	3-AA	388/366	1.02	0.88-1.19	.79
11. rs3741211 (<i>IGF2</i>)	1-TT	1936/1971	N/A	N/A	N/A
11. rs3741211 (<i>IGF2</i>)	2-TC	2367/2269	1.06	0.98-1.16	.17
11. rs3741211 (<i>IGF2</i>)	3-CC	697/760	0.93	0.83-1.05	.28
12. IGF2-05 (<i>IGF2</i>)	1-AA	2651/2694	N/A	N/A	N/A
12. IGF2-05 (<i>IGF2</i>)	2-Aa	1955/1952	1.02	0.94-1.11	.69
12. IGF2-05 (<i>IGF2</i>)	3-aa	394/354	1.13	0.97-1.32	.12
13. IGF2-06 (<i>IGF2</i>)	1-AA	2160/2162	N/A	N/A	N/A
13. IGF2-06 (<i>IGF2</i>)	2-Aa	2237/2284	0.98	0.90-1.07	.66
13. IGF2-06 (<i>IGF2</i>)	3-aa	603/554	1.09	0.96-1.24	.21
14. rs2132571 (<i>IGFBP3</i>)	1-GG	2415/2407	N/A	N/A	N/A
14. rs2132571 (<i>IGFBP3</i>)	2-GA	2163/2157	1	0.92-1.09	.99

SNP ^a (gene)	SNP type	Case number/normal number	OR	95% CI	P value
14. rs2132571 (<i>IGFBP3</i>)	3-AA	422/436	0.97	0.83-1.12	.65
15. rs2471551 (<i>IGFBP3</i>)	1-GG	3225/3284	N/A	N/A	N/A
15. rs2471551 (<i>IGFBP3</i>)	2-GC	1591/1515	1.07	0.98-1.17	.13
15. rs2471551 (<i>IGFBP3</i>)	3-CC	184/201	0.93	0.76-1.15	.54
16. rs2854744 (<i>IGFBP3</i>)	1-AA	1538/1469	N/A	N/A	N/A
16. rs2854744 (<i>IGFBP3</i>)	2-AC	2487/2475	0.96	0.88-1.05	.39
16. rs2854744 (<i>IGFBP3</i>)	3-CC	975/1056	0.88	0.79-0.99	.03
17. rs2132572 (<i>IGFBP3</i>)	1-GG	2908/3027	N/A	N/A	N/A
17. rs2132572 (<i>IGFBP3</i>)	2-GA	1805/1728	1.09	1.00-1.18	.051
17. rs2132572 (<i>IGFBP3</i>)	3-AA	287/245	1.22	1.02-1.46	.03
18. rs3024496 (<i>IL10</i>)	1-TT	1218/1235	N/A	N/A	N/A
18. rs3024496 (<i>IL10</i>)	2-TC	2533/2549	1.01	0.92-1.11	.90
18. rs3024496 (<i>IL10</i>)	3-CC	1249/1216	1.04	0.93-1.17	.49
19. rs1800872 (<i>IL10</i>)	1-CC	3059/3017	N/A	N/A	N/A
19. rs1800872 (<i>IL10</i>)	2-CA	1660/1722	0.95	0.87-1.03	.25
19. rs1800872 (<i>IL10</i>)	3-AA	281/261	1.06	0.89-1.27	.53
20. rs1800890 (<i>IL10</i>)	1-TT	1703/1701	N/A	N/A	N/A
20. rs1800890 (<i>IL10</i>)	2-TA	2455/2508	0.98	0.90-1.07	.63
20. rs1800890 (<i>IL10</i>)	3-AA	842/791	1.06	0.95-1.20	.32
21. rs1554286 (<i>IL10</i>)	1-CC	3400/3446	N/A	N/A	N/A
21. rs1554286 (<i>IL10</i>)	2-CT	1431/1410	1.03	0.94-1.12	.54
21. rs1554286 (<i>IL10</i>)	3-TT	169/144	1.19	0.95-1.49	.15
22. rs1800470 (<i>TGFB1</i>)	1-TT	1850/1914	N/A	N/A	N/A
22. rs1800470 (<i>TGFB1</i>)	2-TC	2372/2399	1.02	0.94-1.11	.62
22. rs1800470 (<i>TGFB1</i>)	3-CC	778/687	1.17	1.04-1.32	.01
23. rs699947 (<i>VEGF</i>)	1-CC	1236/1273	N/A	N/A	N/A
23. rs699947 (<i>VEGF</i>)	2-CA	2511/2463	1.05	0.95-1.16	.33
23. rs699947 (<i>VEGF</i>)	3-AA	1253/1264	1.02	0.91-1.14	.73
24. rs1570360 (<i>VEGF</i>)	1-GG	2278/2341	N/A	N/A	N/A
24. rs1570360 (<i>VEGF</i>)	2-GA	2214/2132	1.07	0.98-1.16	.13
24. rs1570360 (<i>VEGF</i>)	3-AA	508/527	0.99	0.87-1.13	.92
25. rs2010963 (<i>VEGF</i>)	1-GG	2354/2279	N/A	N/A	N/A
25. rs2010963 (<i>VEGF</i>)	2-GC	2133/2157	0.96	0.88-1.04	.31
25. rs2010963 (<i>VEGF</i>)	3-CC	513/564	0.88	0.77-1.01	.07
26. rs3025039 (<i>VEGF</i>)	1-CC	3744/3741	N/A	N/A	N/A
26. rs3025039 (<i>VEGF</i>)	2-CT	1160/1174	0.99	0.90-1.08	.81
26. rs3025039 (<i>VEGF</i>)	3-TT	96/85	1.13	0.84-1.52	.47

^aSNP: single-nucleotide polymorphism.

^bN/A: not applicable.

Comparison of Cases and Controls in Terms of the Effect of a Single SNP

Table 2 compares patients and normal subjects in terms of effect (OR and 95% CI) at a single SNP for growth factor-related genes. Two SNPs within two genes (rs2229765-AA [*IGF1R*] and rs2854744-CC [*IGFBP3*]) showed significant protection associations (rs2229765-AA: OR 0.82, $P=.001$; rs2854744-CC: OR 0.88, $P=.03$) for breast cancer. The minimum and maximum protection associations exhibited ORs of 0.82 and 0.88, respectively, and the other SNPs showed nonsignificant protection associations for breast cancer.

Comparison Between the Proposed HTGA and Existing Algorithms

We compared PSO [34], CPSO [35], and the GA [24] with the HTGA for 2-SNP to 7-SNP barcodes with protection associations (Table 3). ORs (<1) indicate the impact of the protection association of SNP barcodes for the occurrence of breast cancer. A high difference between cases and controls in the SNP barcodes represents informative protection associations, and $P<.05$ indicates a significant difference for the SNP barcode between cases and controls. The identified 3-SNP to 7-SNP barcodes showed that the HTGA provided values with a greater degree of difference as compared with PSO, CPSO, and the GA, indicating that the HTGA identified relevant SNP barcodes with protection associations more effectively (Table 3). HTGA-identified SNP barcodes showed ORs ranging from

0.755 to 0.870 ($P=.003$) for protection associations with breast cancer. The 2-SNP and 3-SNP barcodes in PSO, CPSO, and the GA showed significant differences between cases and controls (2-SNP: $P=.003$, $P=.001$, and $P=.03$, respectively; 3-SNP: $P=.04$, $P=.04$, and $P=.002$, respectively). The 4-SNP barcodes in CPSO and the GA showed significant differences ($P=.04$ and $P=.004$, respectively), and the 5-SNP barcode in the GA also showed a significant difference ($P=.03$). Although CPSO and the GA provided better ORs as compared with the HTGA in all SNP barcodes, the degrees of difference indicated that the SNP barcodes identified by the HTGA were superior to those identified by other methods, and P values $>.05$ indicated that these differences revealed by the models were not significant.

Optimization algorithms have been widely applied to detect relevant high-order SNP barcodes in disease and cancer studies [24,25,34]. Differences between cases and controls are often applied to evaluate the values of SNP barcodes in terms of their fitness function design. As indicated in Table 3, the HTGA effectively identified the relevant protection associations of SNP barcodes for breast cancer. The logistic regression analysis results were strongly validated by the outstanding performance of the HTGA in breast cancer SNP barcode identification. The SNP barcodes detected by the proposed HTGA are simply associations between a barcode and disease, and this type of analysis does not support the inference of causality.

Table 3. Estimation of the best protection single-nucleotide polymorphism barcodes for the occurrence of breast cancer as determined by particle swarm optimization, chaotic particle swarm optimization, the genetic algorithm, and the hybrid Taguchi-genetic algorithm.

Order and method	Combined SNP ^a	SNP genotypes	Control number	Case number	Difference	OR	95% CI	P value
2-SNP								
PSO ^b	1,8	1-3	941	816	125	0.841	0.76-0.93	.001
PSO	N/A ^c	Other	4059	4184	N/A	N/A	N/A	N/A
CPSO ^d	1,8	1-3	941	816	125	0.841	0.76-0.93	.001
CPSO	N/A	Other	4059	4184	N/A	N/A	N/A	N/A
GA ^e	1,10	1-2	1926	1823	103	0.916	0.85-0.99	.03
GA	N/A	Other	3074	3177	N/A	N/A	N/A	N/A
HTGA ^f	10,17	2-1	1309	1179	130 ^g	0.870	0.79-0.95	.003
HTGA	N/A	Other	3691	3821	N/A	N/A	N/A	N/A
3-SNP								
PSO	8,9,22	3-1-2	269	225	44	0.829	0.69-0.99	.043
PSO	N/A	Other	4731	4775	N/A	N/A	N/A	N/A
CPSO	3,8,9	1-3-1	371	319	52	0.850	0.73-0.99	.04
CPSO	N/A	Other	4629	4681	N/A	N/A	N/A	N/A
GA	1,8,15	1-3-1	624	527	97	0.826	0.73-0.94	.002
GA	N/A	Other	4376	4473	N/A	N/A	N/A	N/A
HTGA	1,10,17	1-2-1	1158	1035	123 ^g	0.866	0.79-0.95	.003
HTGA	N/A	Other	3842	3965	N/A	N/A	N/A	N/A
4-SNP								
PSO	4,8,14,22	2-3-1-2	99	76	23	0.764	0.57-1.03	.08
PSO	N/A	Other	4901	4924	N/A	N/A	N/A	N/A
CPSO	10,17,21,23	2-1-1-1	268	223	45	0.824	0.69-0.99	.04
CPSO	N/A	Other	4732	4777	N/A	N/A	N/A	N/A
GA	1,10,17,21	1-2-1-1	795	692	103 ^g	0.850	0.76-0.95	.004
GA	N/A	Other	4205	4308	N/A	N/A	N/A	N/A
HTGA	1,10,17,21	1-2-1-1	795	692	103 ^g	0.850	0.76-0.95	.004
HTGA	N/A	Other	4205	4308	N/A	N/A	N/A	N/A
5-SNP								
PSO	5,6,8,9,26	1-1-3-2-1	91	75	16	0.821	0.60-1.12	.21
PSO	N/A	Other	4909	4925	N/A	N/A	N/A	N/A
CPSO	2,4,8,11,18	1-2-3-1-2	44	32	12	0.726	0.46-1.15	.17
CPSO	N/A	Other	4956	4968	N/A	N/A	N/A	N/A
GA	1,4,15,17,21	1-1-1-1-1	657	585	72	0.876	0.78-0.99	.03
GA	N/A	Other	4343	4415	N/A	N/A	N/A	N/A
HTGA	1,10,17,21,26	1-2-1-1-1	603	520	83 ^g	0.846	0.75-0.96	.009
HTGA	N/A	Other	4397	4480	N/A	N/A	N/A	N/A
6-SNP								
PSO	4,8,15,19,22,24	1-3-2-2-1-3	2	0	2	0.333	0.04-3.20	.34
PSO	N/A	Other	4998	5000	N/A	N/A	N/A	N/A

Order and method	Combined SNP ^a	SNP genotypes	Control number	Case number	Difference	OR	95% CI	P value
CPSO	3,4,12,16,20,24	1-1-1-2-2-3	28	21	7	0.749	0.43-1.32	.32
CPSO	N/A	Other	4972	4979	N/A	N/A	N/A	N/A
GA	1,2,4,6,15,18	1-1-1-1-1-2	276	247	29	0.900	0.75-1.06	.19
GA	N/A	Other	4724	4753	N/A	N/A	N/A	N/A
HTGA	1,10,15,17,21,26	1-2-1-1-1-1	394	327	67 ^g	0.818	0.70-0.95	.01
HTGA	N/A	Other	4606	4673	N/A	N/A	N/A	N/A
7-SNP								
PSO	5,8,11,13,14,24,25	1-1-3-1-1-2-1	6	3	3	0.500	0.13-2.00	.33
PSO	N/A	Other	4994	4997	N/A	N/A	N/A	N/A
CPSO	10,12,16,17,19,22,26	2-2-2-1-2-2-1	27	20	7	0.740	0.41-1.32	.31
CPSO	N/A	Other	4973	4980	N/A	N/A	N/A	N/A
GA	1,2,6,7,10,14,15	1-1-1-3-2-1-1	38	25	13	0.656	0.40-1.09	.10
GA	N/A	Other	4962	4975	N/A	N/A	N/A	N/A
HTGA	1,10,13,15,17,21,26	1-2-2-1-1-1-1	185	141	44 ^g	0.755	0.60-0.94	.01
HTGA	N/A	Other	4815	4859	N/A	N/A	N/A	N/A

^aSNP: single-nucleotide polymorphism.

^bPSO: particle swarm optimization.

^cN/A: not applicable.

^dCPSO: chaotic particle swarm optimization.

^eGA: genetic algorithm.

^fHTGA: hybrid Taguchi-genetic algorithm.

^gThe best results in the *n*-SNP barcodes.

Discussion

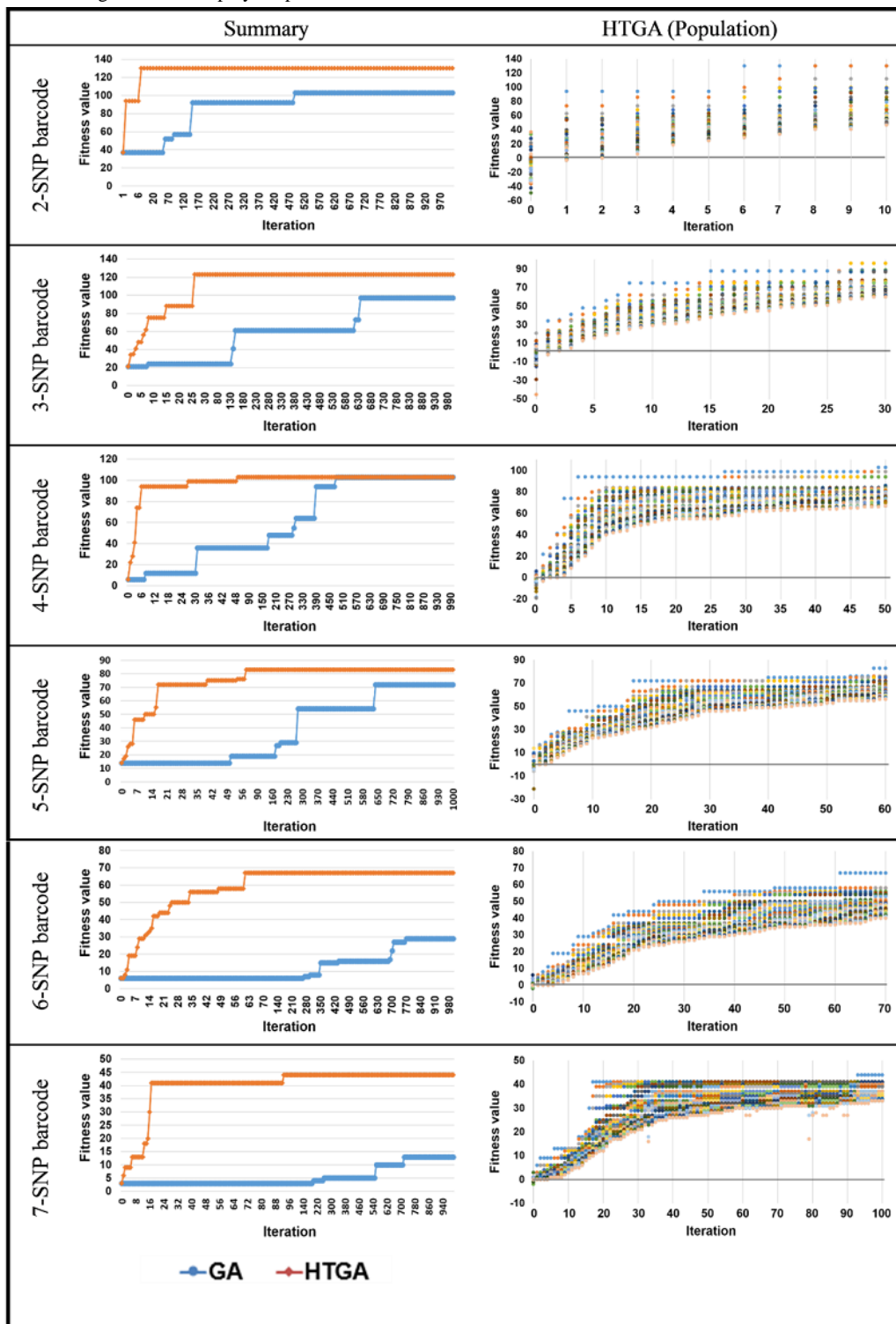
Principal Findings

Many breast cancer studies have identified the associations among the effects of important related genes [36-42], including genes related to DNA repair [43,44] and estrogen-response genes [45]. In this study, we introduced a HTGA to identify the SNP barcodes among 26 breast cancer-related SNPs. The HTGA-generated SNP barcodes were examined to determine their protective effects against breast cancer. The results suggest that nonrelevant SNPs might cumulatively reduce the risk of breast cancer, as indicated by the HTGA-generated preventive SNP barcodes. A search space consisting of SNP barcode combinations can generate numerous local optima in multiple regions. These local optima raise challenges for optimization algorithm search operations, because the heuristic and stochastic properties of such optimization algorithms can easily cause searches to become trapped in local optima. A GA population can be updated by referring to other chromosomes to determine the next position in the search space. However, GA operations can result in stagnation if the chromosomes are similar; points of stagnation in a search space are referred to as local optima. The computational processes and comparisons are shown in Figure 2. A Taguchi system is a nonlinear system with deterministic dynamic behavior owing to its ergodic and stochastic properties. Taguchi methods are used to enhance GA crossover operations, and these methods can be remarkably

helpful for avoiding population entrapment in local optima because improved solutions can be found through experimentation. Because the population learns from experience, it can be said to exhibit population intelligence. The HTGA can converge quickly to excellent fitness values for SNP barcodes, whereas the GA is very slow to converge and the results are worse than those of the HTGA (Figure 2), indicating that the GA can very easily result in stagnation in regions that may not include any global optima. However, the population is effectively improved in the HTGA, and Figure 2 shows that the fitness values of chromosomes clearly increase over time, proving that the proposed Taguchi method can be used to improve GA performance to identify SNP barcodes. Moreover, our results prove the ability of this Taguchi-based GA to solve SNP barcode identification problems. The optimal parameters of the HTGA could be further analyzed for enhancing the detection ability of SNP barcodes. Our HTGA includes the probability of crossover and mutation. A further investigation with more data sets is required to determine the optimal parameters. Moreover, selection, crossover, mutation, and replacement operations can be analyzed to determine the superior operation strategy for enhancing the ability of our HTGA to detect potential SNP barcodes. If the HTGA is applied for clinical data, we suggest considering permutation testing to examine the relevance of the results obtained. For each trial in permutation testing, the case/control labels would be scrambled, and the algorithm would then search for an optimal solution.

After numerous trials, we would be able to determine the number of times a solution at least as good as the one from the original data is found and then determine if the algorithm is simply fitting the data or identifying underlying associations.

Figure 2. Comparison of improvements to fitness values between the genetic algorithm (GA) and hybrid Taguchi-genetic algorithm (HTGA).The images on the left compare GA and HTGA search results for 1000 iterations. The images on the right present the fitness values of an HTGA population at specific iterations. SNP: single-nucleotide polymorphism.



Conclusions

An HTGA was proposed to effectively identify relevant SNP barcodes among genes related to breast cancer. The study results demonstrated that the HTGA could effectively detect SNP

barcodes for problems with numerous high-order SNP barcode combinations. The proposed Taguchi method can improve the GA via the identification of high-dimensional SNP barcodes, and hence, it is integrated following GA crossover operations to systematically optimize chromosomes and thus enhance their

values. Moreover, the HTGA can effectively converge to a promising region within the problem space and provide excellent SNP barcode identification. In this study, large data sets were used to evaluate and compare the performances of the GA, PSO,

CPSO, and the HTGA, and the results indicated that the HTGA can effectively identify relevant high-order SNP barcodes in breast cancer.

Acknowledgments

This work was supported by the National Science Council, Taiwan (108-2221-E-214-019-MY3 and 108-2221-E-992-031-MY3).

Conflicts of Interest

None declared.

References

1. Sharma R. Breast cancer incidence, mortality and mortality-to-incidence ratio (MIR) are associated with human development, 1990-2016: evidence from Global Burden of Disease Study 2016. *Breast Cancer* 2019 Jul;26(4):428-445. [doi: [10.1007/s12282-018-00941-4](https://doi.org/10.1007/s12282-018-00941-4)] [Medline: [30604398](https://pubmed.ncbi.nlm.nih.gov/30604398/)]
2. Liu F, Lin H, Kuo C, See L, Chiou M, Yu H. Epidemiology and survival outcome of breast cancer in a nationwide study. *Oncotarget* 2017 Mar 07;8(10):16939-16950 [FREE Full text] [doi: [10.18632/oncotarget.15207](https://doi.org/10.18632/oncotarget.15207)] [Medline: [28199975](https://pubmed.ncbi.nlm.nih.gov/28199975/)]
3. Abubakar M, Sung H, Bcr D, Guida J, Tang TS, Pfeiffer RM, et al. Breast cancer risk factors, survival and recurrence, and tumor molecular subtype: analysis of 3012 women from an indigenous Asian population. *Breast Cancer Res* 2018 Sep 18;20(1):114 [FREE Full text] [doi: [10.1186/s13058-018-1033-8](https://doi.org/10.1186/s13058-018-1033-8)] [Medline: [30227867](https://pubmed.ncbi.nlm.nih.gov/30227867/)]
4. Visser LL, Elshof LE, Schaapveld M, van de Vijver K, Groen EJ, Almekinders MM, et al. Clinicopathological Risk Factors for an Invasive Breast Cancer Recurrence after Ductal Carcinoma —A Nested Case–Control Study. *Clin Cancer Res* 2018 Apr 23;24(15):3593-3601. [doi: [10.1158/1078-0432.ccr-18-0201](https://doi.org/10.1158/1078-0432.ccr-18-0201)]
5. Park S, Han W, Kim J, Kim MK, Lee E, Yoo T, et al. Risk Factors Associated with Distant Metastasis and Survival Outcomes in Breast Cancer Patients with Locoregional Recurrence. *J Breast Cancer* 2015 Jun;18(2):160-166 [FREE Full text] [doi: [10.4048/jbc.2015.18.2.160](https://doi.org/10.4048/jbc.2015.18.2.160)] [Medline: [26155292](https://pubmed.ncbi.nlm.nih.gov/26155292/)]
6. Marsaux CF, Celis-Morales C, Livingstone KM, Fallaize R, Kolossa S, Hallmann J, et al. Changes in Physical Activity Following a Genetic-Based Internet-Delivered Personalized Intervention: Randomized Controlled Trial (Food4Me). *J Med Internet Res* 2016 Feb 05;18(2):e30 [FREE Full text] [doi: [10.2196/jmir.5198](https://doi.org/10.2196/jmir.5198)] [Medline: [26851191](https://pubmed.ncbi.nlm.nih.gov/26851191/)]
7. Shin SJ, You SC, Park YR, Roh J, Kim J, Haam S, et al. Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. *J Med Internet Res* 2019 Mar 26;21(3):e13249 [FREE Full text] [doi: [10.2196/13249](https://doi.org/10.2196/13249)] [Medline: [30912749](https://pubmed.ncbi.nlm.nih.gov/30912749/)]
8. Yang C, Chuang L, Lin Y. Epistasis Analysis Using an Improved Fuzzy C-Means-Based Entropy Approach. *IEEE Trans. Fuzzy Syst* 2020 Apr;28(4):718-730. [doi: [10.1109/tfuzz.2019.2914629](https://doi.org/10.1109/tfuzz.2019.2914629)]
9. Yang C, Moi S, Ou-Yang F, Chuang L, Hou M, Lin Y. Identifying Risk Stratification Associated With a Cancer for Overall Survival by Deep Learning-Based CoxPH. *IEEE Access* 2019;7:67708-67717. [doi: [10.1109/access.2019.2916586](https://doi.org/10.1109/access.2019.2916586)]
10. Moi S, Lee Y, Chuang L, Yuan SF, Ou-Yang F, Hou M, et al. Cumulative receiver operating characteristics for analyzing interaction between tissue visfatin and clinicopathologic factors in breast cancer progression. *Cancer Cell Int* 2018;18:19 [FREE Full text] [doi: [10.1186/s12935-018-0517-z](https://doi.org/10.1186/s12935-018-0517-z)] [Medline: [29449787](https://pubmed.ncbi.nlm.nih.gov/29449787/)]
11. Kotredes KP, Razmpour R, Lutton E, Alfonso-Prieto M, Ramirez SH, Gamero AM. Characterization of cancer-associated IDH2 mutations that differ in tumorigenicity, chemosensitivity and 2-hydroxyglutarate production. *Oncotarget* 2019 Apr 12;10(28):2675-2692 [FREE Full text] [doi: [10.18632/oncotarget.26848](https://doi.org/10.18632/oncotarget.26848)] [Medline: [31105869](https://pubmed.ncbi.nlm.nih.gov/31105869/)]
12. Šolman M, Ligabue A, Blaževič O, Jaiswal A, Zhou Y, Liang H, et al. Specific cancer-associated mutations in the switch III region of Ras increase tumorigenicity by nanocluster augmentation. *Elife* 2015 Aug 14;4:e08905 [FREE Full text] [doi: [10.7554/eLife.08905](https://doi.org/10.7554/eLife.08905)] [Medline: [26274561](https://pubmed.ncbi.nlm.nih.gov/26274561/)]
13. Derouet M, Wu X, May L, Hoon Yoo B, Sasazuki T, Shirasawa S, et al. Acquisition of anoikis resistance promotes the emergence of oncogenic K-ras mutations in colorectal cancer cells and stimulates their tumorigenicity in vivo. *Neoplasia* 2007 Jul;9(7):536-545 [FREE Full text] [doi: [10.1593/neo.07217](https://doi.org/10.1593/neo.07217)] [Medline: [17710156](https://pubmed.ncbi.nlm.nih.gov/17710156/)]
14. Petros JA, Baumann AK, Ruiz-Pesini E, Amin MB, Sun CQ, Hall J, et al. mtDNA mutations increase tumorigenicity in prostate cancer. *Proc Natl Acad Sci U S A* 2005 Jan 18;102(3):719-724 [FREE Full text] [doi: [10.1073/pnas.0408894102](https://doi.org/10.1073/pnas.0408894102)] [Medline: [15647368](https://pubmed.ncbi.nlm.nih.gov/15647368/)]
15. Weich N, Ferri C, Moiraghi B, Bengió R, Giere I, Pavlovsky C, et al. GSTM1 and GSTP1, but not GSTT1 genetic polymorphisms are associated with chronic myeloid leukemia risk and treatment response. *Cancer Epidemiol* 2016 Oct;44:16-21. [doi: [10.1016/j.canep.2016.07.008](https://doi.org/10.1016/j.canep.2016.07.008)] [Medline: [27454607](https://pubmed.ncbi.nlm.nih.gov/27454607/)]
16. Fu OY, Chang H, Lin Y, Chuang L, Hou M, Yang C. Breast cancer-associated high-order SNP-SNP interaction of CXCL12/CXCR4-related genes by an improved multifactor dimensionality reduction (MDR-ER). *Oncol Rep* 2016 Sep;36(3):1739-1747. [doi: [10.3892/or.2016.4956](https://doi.org/10.3892/or.2016.4956)] [Medline: [27461876](https://pubmed.ncbi.nlm.nih.gov/27461876/)]

17. Tang J, Chuang L, Hsi E, Lin Y, Yang C, Chang H. Identifying the association rules between clinicopathologic factors and higher survival performance in operation-centric oral cancer patients using the Apriori algorithm. *Biomed Res Int* 2013;2013:359634 [FREE Full text] [doi: [10.1155/2013/359634](https://doi.org/10.1155/2013/359634)] [Medline: [23984353](https://pubmed.ncbi.nlm.nih.gov/23984353/)]
18. Yang C, Wu K, Dahms H, Chuang L, Chang H. Single nucleotide polymorphism barcoding of cytochrome c oxidase I sequences for discriminating 17 species of Columbidae by decision tree algorithm. *Ecol Evol* 2017 Jul;7(13):4717-4725 [FREE Full text] [doi: [10.1002/ece3.3045](https://doi.org/10.1002/ece3.3045)] [Medline: [28690801](https://pubmed.ncbi.nlm.nih.gov/28690801/)]
19. Yang C, Lin Y, Chuang L. Class Balanced Multifactor Dimensionality Reduction to Detect Gene–Gene Interactions. *IEEE/ACM Trans. Comput. Biol. and Bioinf* 2020 Jan 1;17(1):71-81. [doi: [10.1109/tcbb.2018.2858776](https://doi.org/10.1109/tcbb.2018.2858776)]
20. Ou-Yang F, Lin Y, Chuang L, Chang H, Yang C, Hou M. The Combinational Polymorphisms of ORAI1 Gene Are Associated with Preventive Models of Breast Cancer in the Taiwanese. *Biomed Res Int* 2015;2015:281263 [FREE Full text] [doi: [10.1155/2015/281263](https://doi.org/10.1155/2015/281263)] [Medline: [26380267](https://pubmed.ncbi.nlm.nih.gov/26380267/)]
21. Yang P, Ho JW, Yang YH, Zhou BB. Gene-gene interaction filtering with ensemble of filters. *BMC Bioinformatics* 2011 Feb 15;12(S1). [doi: [10.1186/1471-2105-12-s1-s10](https://doi.org/10.1186/1471-2105-12-s1-s10)]
22. Chuang L, Lane H, Lin Y, Lin M, Yang C, Chang H. Identification of SNP barcode biomarkers for genes associated with facial emotion perception using particle swarm optimization algorithm. *Ann Gen Psychiatry* 2014;13(1):15. [doi: [10.1186/1744-859x-13-15](https://doi.org/10.1186/1744-859x-13-15)]
23. Yan R, Cao J, Song C, Chen Y, Wu Z, Wang K, et al. Polymorphisms in lncRNA HOTAIR and susceptibility to breast cancer in a Chinese population. *Cancer Epidemiol* 2015 Dec;39(6):978-985. [doi: [10.1016/j.canep.2015.10.025](https://doi.org/10.1016/j.canep.2015.10.025)] [Medline: [26547792](https://pubmed.ncbi.nlm.nih.gov/26547792/)]
24. Chang W, Fang Y, Chang H, Chuang L, Lin Y, Hou M, et al. Identifying association model for single-nucleotide polymorphisms of ORAI1 gene for breast cancer. *Cancer Cell Int* 2014 Mar 31;14(1):29 [FREE Full text] [doi: [10.1186/1475-2867-14-29](https://doi.org/10.1186/1475-2867-14-29)] [Medline: [24685237](https://pubmed.ncbi.nlm.nih.gov/24685237/)]
25. Chen J, Chuang L, Lin Y, Liou C, Lin T, Lee W, et al. Genetic algorithm-generated SNP barcodes of the mitochondrial D-loop for chronic dialysis susceptibility. *Mitochondrial DNA* 2014 Jun;25(3):231-237. [doi: [10.3109/19401736.2013.796513](https://doi.org/10.3109/19401736.2013.796513)] [Medline: [23777414](https://pubmed.ncbi.nlm.nih.gov/23777414/)]
26. Yang C, Moi S, Lin Y, Chuang L. Genetic Algorithm Combined with a Local Search Method for Identifying Susceptibility Genes. *Journal of Artificial Intelligence and Soft Computing Research* 2016;6(3):203-212. [doi: [10.1515/jaiscr-2016-0015](https://doi.org/10.1515/jaiscr-2016-0015)]
27. Holland J. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, Massachusetts: MIT Press; 1992.
28. Bendell A, Disney J, Pridmore W. *Taguchi Methods: Applications in World Industry*. Berlin, Germany: Springer Verlag; 1989.
29. Miller B, Goldberg D. Genetic algorithms, tournament selection, and the effects of noise. *Complex Systems* 1995;9(3):193-212.
30. Eberhart R, Kennedy J. A new optimizer using particle swarm theory. In: *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*. 1995 Presented at: Sixth International Symposium on Micro Machine and Human Science; October 4-6, 1995; Nagoya, Japan p. 39-43. [doi: [10.1109/mhs.1995.494215](https://doi.org/10.1109/mhs.1995.494215)]
31. Chuang L, Chang H, Lin M, Yang C. Chaotic particle swarm optimization for detecting SNP–SNP interactions for CXCL12-related genes in breast cancer prevention. *European Journal of Cancer Prevention* 2012;21(4):336-342. [doi: [10.1097/cej.0b013e32834e31f6](https://doi.org/10.1097/cej.0b013e32834e31f6)]
32. Mechanic LE, Luke BT, Goodman JE, Chanock SJ, Harris CC. Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions. *BMC Bioinformatics* 2008 Mar 06;9:146 [FREE Full text] [doi: [10.1186/1471-2105-9-146](https://doi.org/10.1186/1471-2105-9-146)] [Medline: [18325117](https://pubmed.ncbi.nlm.nih.gov/18325117/)]
33. Pharoah PD, Tyrer J, Dunning AM, Easton DF, Ponder BA, SEARCH Investigators. Association between common variation in 120 candidate genes and breast cancer risk. *PLoS Genet* 2007 Mar 16;3(3):e42 [FREE Full text] [doi: [10.1371/journal.pgen.0030042](https://doi.org/10.1371/journal.pgen.0030042)] [Medline: [17367212](https://pubmed.ncbi.nlm.nih.gov/17367212/)]
34. Wu S, Chuang L, Lin Y, Ho W, Chiang F, Yang C, et al. Particle swarm optimization algorithm for analyzing SNP-SNP interaction of renin-angiotensin system genes against hypertension. *Mol Biol Rep* 2013 Jul;40(7):4227-4233. [doi: [10.1007/s11033-013-2504-8](https://doi.org/10.1007/s11033-013-2504-8)] [Medline: [23695493](https://pubmed.ncbi.nlm.nih.gov/23695493/)]
35. Chuang L, Chang H, Lin M, Yang C. Chaotic particle swarm optimization for detecting SNP–SNP interactions for CXCL12-related genes in breast cancer prevention. *European Journal of Cancer Prevention* 2012;21(4):336-342. [doi: [10.1097/cej.0b013e32834e31f6](https://doi.org/10.1097/cej.0b013e32834e31f6)]
36. Chen L, Li W, Zhang L, Wang H, He W, Tai J, et al. Disease gene interaction pathways: a potential framework for how disease genes associate by disease-risk modules. *PLoS One* 2011;6(9):e24495 [FREE Full text] [doi: [10.1371/journal.pone.0024495](https://doi.org/10.1371/journal.pone.0024495)] [Medline: [21915342](https://pubmed.ncbi.nlm.nih.gov/21915342/)]
37. Yin J, Lu K, Lin J, Wu L, Hildebrandt MA, Chang DW, et al. Genetic variants in TGF- β pathway are associated with ovarian cancer risk. *PLoS One* 2011;6(9):e25559 [FREE Full text] [doi: [10.1371/journal.pone.0025559](https://doi.org/10.1371/journal.pone.0025559)] [Medline: [21984931](https://pubmed.ncbi.nlm.nih.gov/21984931/)]
38. Ricceri F, Guarrera S, Sacerdote C, Polidoro S, Allione A, Fontana D, et al. ERCC1 haplotypes modify bladder cancer risk: a case-control study. *DNA Repair (Amst)* 2010 Feb 04;9(2):191-200. [doi: [10.1016/j.dnarep.2009.12.002](https://doi.org/10.1016/j.dnarep.2009.12.002)] [Medline: [20061190](https://pubmed.ncbi.nlm.nih.gov/20061190/)]

39. Cauchi S, Meyre D, Durand E, Proença C, Marre M, Hadjadj S, et al. Post genome-wide association studies of novel genes associated with type 2 diabetes show gene-gene interaction and high predictive value. *PLoS One* 2008 May 07;3(5):e2031 [FREE Full text] [doi: [10.1371/journal.pone.0002031](https://doi.org/10.1371/journal.pone.0002031)] [Medline: [18461161](https://pubmed.ncbi.nlm.nih.gov/18461161/)]
40. Lin G, Tseng H, Chang C, Chuang L, Liu C, Yang C, et al. SNP combinations in chromosome-wide genes are associated with bone mineral density in Taiwanese women. *Chin J Physiol* 2008 Feb 29;51(1):32-41. [Medline: [18551993](https://pubmed.ncbi.nlm.nih.gov/18551993/)]
41. Yang C, Lin Y, Yen C, Chuang L, Chang H. A systematic gene-gene and gene-environment interaction analysis of DNA repair genes XRCC1, XRCC2, XRCC3, XRCC4, and oral cancer risk. *OMICS* 2015 Apr;19(4):238-247. [doi: [10.1089/omi.2014.0121](https://doi.org/10.1089/omi.2014.0121)] [Medline: [25831063](https://pubmed.ncbi.nlm.nih.gov/25831063/)]
42. Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, et al. Cumulative Association of Five Genetic Variants with Prostate Cancer. *N Engl J Med* 2008 Feb 28;358(9):910-919. [doi: [10.1056/nejmoa075819](https://doi.org/10.1056/nejmoa075819)]
43. Conde J, Silva SN, Azevedo AP, Teixeira V, Pina JE, Rueff J, et al. Association of common variants in mismatch repair genes and breast cancer susceptibility: a multigene study. *BMC Cancer* 2009 Sep 25;9:344 [FREE Full text] [doi: [10.1186/1471-2407-9-344](https://doi.org/10.1186/1471-2407-9-344)] [Medline: [19781088](https://pubmed.ncbi.nlm.nih.gov/19781088/)]
44. Han W, Kim K, Yang S, Noh D, Kang D, Kwack K. SNP-SNP interactions between DNA repair genes were associated with breast cancer risk in a Korean population. *Cancer* 2012 Feb 01;118(3):594-602 [FREE Full text] [doi: [10.1002/cncr.26220](https://doi.org/10.1002/cncr.26220)] [Medline: [21751184](https://pubmed.ncbi.nlm.nih.gov/21751184/)]
45. Yu J, Hsiung C, Hsu H, Bao B, Chen S, Hsu G, et al. Genetic variation in the genome-wide predicted estrogen response element-related sequences is associated with breast cancer development. *Breast Cancer Res* 2011 Jan 31;13(1):R13 [FREE Full text] [doi: [10.1186/bcr2821](https://doi.org/10.1186/bcr2821)] [Medline: [21281495](https://pubmed.ncbi.nlm.nih.gov/21281495/)]

Abbreviations

- CPSO:** chaotic particle swarm optimization
GA: genetic algorithm
HTGA: hybrid Taguchi-genetic algorithm
OA: orthogonal array
PSO: particle swarm optimization
SNP: single-nucleotide polymorphism
SNR: signal-to-noise ratio

Edited by C Lovis; submitted 04.11.19; peer-reviewed by BTL Brian T Luke, HW Chang; comments to author 15.12.19; revised version received 09.02.20; accepted 08.04.20; published 17.06.20

Please cite as:

Chuang LY, Yang CS, Yang HS, Yang CH

Identification of High-Order Single-Nucleotide Polymorphism Barcodes in Breast Cancer Using a Hybrid Taguchi-Genetic Algorithm: Case-Control Study

JMIR Med Inform 2020;8(6):e16886

URL: <https://medinform.jmir.org/2020/6/e16886>

doi: [10.2196/16886](https://doi.org/10.2196/16886)

PMID: [32554381](https://pubmed.ncbi.nlm.nih.gov/32554381/)

©Li-Yeh Chuang, Cheng-San Yang, Huai-Shuo Yang, Cheng-Hong Yang. Originally published in *JMIR Medical Informatics* (<http://medinform.jmir.org>), 17.06.2020. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Medical Informatics*, is properly cited. The complete bibliographic information, a link to the original publication on <http://medinform.jmir.org/>, as well as this copyright and license information must be included.