Complex & Intelligent Systems (2023) 9:637–658
https://doi.org/10.1007/s40747-022-00813-7

ORIGINAL ARTICLE

MTHSA-DHEI: multitasking harmony search algorithm for detecting high-order SNP epistatic interactions

Shouheng Tuo[1,2,3] · Chao Li[1,2,3] · Fan Liu[1,2,3] · Aimin Li[4] · Lang He[1,2,3] · Zong Woo Geem[5] · JunLiang Shang[6] · Haiyan Liu[1,2,3] · YanLing Zhu[1,2,3] · ZengYu Feng[1,2,3] · TianRui Chen[1,2,3]

Received: 12 January 2022 / Accepted: 16 June 2022 / Published online: 27 July 2022

Abstract

Genome-wide association studies have succeeded in identifying genetic variants associated with complex diseases, but the findings have not been well interpreted biologically. Although it is widely accepted that epistatic interactions of high-*order* single nucleotide polymorphisms (SNPs) [(1) ***Single nucleotide polymorphisms (SNP)*** are mainly deoxyribonucleic acid (DNA) sequence polymorphisms caused by variants at a single nucleotide at the genome level. They are the most common type of heritable variation in humans.] are important causes of complex diseases, the combinatorial explosion of millions of SNPs and multiple tests impose a large computational burden. Moreover, it is extremely challenging to correctly distinguish high-*order* SNP epistatic interactions from other high-*order* SNP combinations due to small sample sizes. In this study, a multitasking harmony search algorithm (MTHSA-DHEI) is proposed for detecting high-*order* epistatic interactions [(2) In classical genetics, if genes X1 and X2 are mutated and each mutation by itself produces a unique disease status (phenotype) but the mutations together cause the same disease status as the gene X1 mutation, gene X1 is ***epistatic*** and gene X2 is hypostatic, and gene X1 has an epistatic effect (main effect) on disease status. In this work, a high-order ***epistatic interaction*** occurs when two or more SNP loci have a joint influence on disease status.], with the goal of simultaneously detecting multiple types of high-*order* (k_1-*order*, k_2-*order*, …, k_n-*order*) SNP epistatic interactions. Unified coding is adopted for multiple tasks, and four complementary association evaluation functions are employed to improve the capability of discriminating the high-*order* SNP epistatic interactions. We compare the proposed MTHSA-DHEI method with four excellent methods for detecting high-*order* SNP interactions for 8 high-*order* epistatic interaction models with no marginal effect (EINMEs) and 12 epistatic interaction models with marginal effects (EIMEs) (*) and implement the MTHSA-DHEI algorithm with a real dataset: age-related macular degeneration (AMD). The experimental results indicate that MTHSA-DHEI has power and an F1-score exceeding 90% for all EIMEs and five EINMEs and reduces the computational time by more than 90%. It can efficiently perform multiple high-*order* detection tasks for high-*order* epistatic interactions and improve the discrimination ability for diverse epistasis models.

Keywords Multitasking · Harmony search algorithm · Single nucleotide polymorphisms · Epistatic interaction

Introduction

Genome-wide association studies (GWASs) are widely considered one of the most promising technologies for elucidating complex relationships between genotype and phenotype

✉ Shouheng Tuo
tuo_sh@126.com; tuo_sh@xupt.edu.cn

1 School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China

2 Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an 710121, Shaanxi, China

3 Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an 710121, Shaanxi, China

4 School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, Shaanxi, China

5 Department of Energy IT, Gachon University, Seongnam 13120, Korea

6 School of Computer Science, Qufu Normal University, Rizhao 276826, China

[1], such as the [1]causes of complex diseases, due to the rapid development of high-throughput sequencing technology and dramatic declines in sequencing costs. GWASs are dedicated to detecting genetic variants associated with complex traits/diseases from single nucleotide polymorphisms (SNPs), which are the most common genetic variations in human deoxyribonucleic acid (DNA) sequences [2–5].

Many important and interesting findings have been made by GWASs using single-SNP-based and SNP-pair-based methods. Single-SNP analysis approaches for GWASs, such as the single-SNP test [5], compare the relative frequencies of genotypes between case and control samples independently of other SNP loci, and some results have been successfully translated to candidate drugs [6, 7]. Nevertheless, most studies fail to effectively explain the causal SNPs of complex diseases. One important reason is that most studies focus on discovering the contribution of single SNPs to complex disease status/traits in isolation, and SNPs with a small effect on the phenotype were neglected in further analysis [8]. In recent years, various multi-SNP methods have been employed for GWASs, such as penalized regression [9–11], which can eliminate SNPs associated only with the phenotype due to their linkage disequilibrium (LD) with causal SNPs [11]. An increasing number of studies indicate that epistatic interactions across the whole genome ubiquitously exist in relation to complex diseases [12]. Epistatic interaction generally refers to joint interaction effects among multiple genetic variants in the genome, where the effect of a set of genes or SNPs on a phenotype is unequal to the sum of their independent contributions [11]. Epistatic interactions are now widely regarded to determine individual susceptibility to complex diseases [8, 13].

Detecting high-*order* epistatic interactions in the human genome has become a very important goal in GWASs, but it is also extremely challenging because there are hundreds of thousands of SNPs in the human genome, creating a very complex "combination explosion problem". Current computers are not capable of determining whether each *k*th-*order* ($k > 2$) SNP combination has an epistatic interaction effect in a limited time. To address this problem, high-performance computing and heuristic searches have been presented to accelerate the detection of high-*order* epistatic interactions. High-performance computing usually adopts graphics processing units (GPUs) and parallel processing techniques to improve the speed of computers. Guo et al. employed cloud computing to detect high-*order* epistatic interactions [14], and forty virtual machines were adopted to accelerate the detection of such interactions. Yang et al. [15] developed a GPU-based permutation tool to accelerate the detection of SNP-SNP interactions based on the likelihood ratio (LR) test with the assumption that the statistic follows a χ^2 distribution. Cecilia et al. [16] presented a tool called MPI3SNP that implements a multicentral processing unit (CPU) and multi-GPU clusters to detect 3rd-order epistatic interactions. Alex Upton et al. [7] reviewed high-performance computing and cloud computing used to detect epistasis in detail and dissected different computational approaches to analyse epistatic interactions in disease-related genetic datasets. GPU and parallel processing techniques can speed up detection but are insufficient for high-order (> 3) epistatic interactions because the time complexity of detecting high-order SNP epistatic interactions is not reduced if the search algorithm still has high time complexity (i.e., exhaustive search algorithm).

To reduce the computational burden, heuristic search techniques, such as the Monte Carlo method [13, 17, 18], the spanning tree method [19], and swarm intelligence search algorithms (SISAs) [20, 21], use current information about the target problem as heuristic information that can improve search efficiency and reduce the number of searches. The Monte Carlo method employs random sampling procedures to explore potential SNP epistatic interactions and can speed up epistasis detection, but its power is often unsatisfactory. Zhang, Y et al. presented a Bayesian partition model (called Beam) for detecting SNP epistatic interactions, and they employed Markov chain Monte Carlo (MCMC) sampling to compute the posterior probability of SNP markers [13]. Beam models have a very rapid search speed but easily miss epistatic interactions with weak marginal effects on disease status. Wang W introduced a minimum spanning tree structure for exhaustively detecting two-locus epistasis [19]. The minimum spanning tree-based method is powerful for detecting 2nd-order epistatic interactions with marginal effects but largely inefficient for the detection of high-order SNP epistatic interactions with weak marginal effects. Shanwen Sun et al. analysed statistical modelling and machine learning approaches for identifying SNP epistatic interactions in detail [11].

Due to the powerful exploration capability of a high-dimensional search space, the SISA has received much attention in recent years for high-*order* epistatic interaction detection. Moore JH et al. employed a genetic algorithm (GA) to discover complex genetic models [20, 21] and presented a grid-based stochastic search algorithm (named Crush-MDR) [22], which adopts genetic model-free multifactor dimensionality reduction (MDR) to calculate the associations between SNP combinations and disease status in order to accelerate the detection of high-order epistatic interactions. Crush-MDR reduces the time complexity of the search process, but the objective function MDR is

[1] (*) In EIMEs, one or more SNP loci also have a marginal effect on disease status, resulting in an epistatic interaction model with additive effects. In EINMEs, each SNP locus has no or a very weak marginal effect on disease status.

computationally expensive. Wang et al. proposed a two-stage ant colony optimization (ACO) algorithm (named AntEpiSeeker) to detect epistatic interactions [23], which employs the chi-square (χ^2) test to evaluate the scores of SNP combinations. In the 1st stage of AntEpiSeeker, ACO is used to select suspected SNP sets with high χ^2 scores, and the 2nd stage of AntEpiSeeker conducts an exhaustive search with the suspected SNPs. Shang and Sun et al. conducted an in-depth study on gene–gene interactions via ACO [24–26], and their research concentrated on the identification of epistatic models and the improvement of ACO. Differential evolution-based methods were adopted by Yang et al. [27, 28] to detect epistatic interactions, and improved MDR was used to measure associations. Aflakparast, M. et al. introduced a cuckoo search epistasis (CSE) detection algorithm to identify high-order SNP epistatic interactions in which each single SNP has a small effect on disease status. The CSE detection algorithm first divides all SNP loci into M groups based on their relevance, and kth-order SNP combinations are then chosen from the M groups [29].

Tuo et al. proposed three harmony search (HS)-based epistatic detection algorithms (FHSA-SED [30], NHSA-DHSC [31], and MP-HS-DHSI [32]) because of the performance advantages of HS, such as its powerful exploration ability and fast speed. HS is a very simple optimization algorithm that has shown outstanding performance in solving both combinational optimization problems and real number optimization problems. FHSA-SED aims to discover SNP-pair (2nd-*order*) interactions using HS and two scoring functions (the Gini index and Bayesian network-based K2 score) to evaluate the associations between SNP pairs and disease status. NHSA-DHSC presents a niche HS for detecting high-*order* epistatic interactions, in which a niche strategy is used to record local optimal solutions and avoid repeated searches in local regions. The MP-HS-DHSI algorithm employs multipopulational and multiple scoring functions to improve the exploration power of HS and overcome the preference for disease models. In terms of search speed and detection power, it outperforms FHSA-SED and NHSA-DHSC in detecting high-order SNP epistatic interactions, but for an unknown detection task, it also requires the detection of 2nd-order, 3rd-order, …, Kth-order epistatic interactions one by one.

Although the SISA has made some progress in accelerating the detection of high-*order* epistatic interactions, it still faces two challenges:

Search. Finding *k*th-*order* (a combination of *k* SNP loci) epistatic interactions among over hundreds of thousands of SNPs in the whole genome in a limited time is very difficult due to the large number of SNP combinations, which is a complex combination explosion problem. For example, the number of 3rd-*order* SNP combinations for 1,000,000 SNPs is larger than 1.6667×10^{17}. In particular, if the SNPs in high-*order* epistatic interactions have very weak or no marginal effect on disease status/complex traits, the SISA is inefficient or nearly powerless because there are no valid clues to guide the population to locate the causal SNP epistatic interaction among the extreme number of SNP combinations.

Discrimination. The discriminating function (objective function) adopted to calculate the associations between SNP combinations and phenotypes is crucial for the SISA. Faced with such a large number of SNP combinations, discrimination functions with light computational requirements should be considered first for SISAs. Bayesian network-based methods [33–35], Shannon entropy-based methods (i.e., mutual information and conditional entropy) [36] and statistical test methods (i.e., chi-square tests [37]) are lightweight methods that have been widely used to evaluate associations, but none of them are considered effective for all types of epistatic interaction models. Machine learning approaches, such as MDR [38–40], random forest and neural networks, are statistical-free methods with strong applicability for evaluating various disease models, but the high computational burden limits the usefulness of these methods as objective functions of SISAs.

To address the above challenges, this study aims to improve performance in detecting high-order SNP epistatic interactions in the following two aspects:

(1) A multitasking HS algorithm with three stages is developed to improve detection speed and power.
(2) Four complementary association evaluation functions are employed to improve the discrimination ability of various disease models.

To the best of our knowledge, the existing SISAs for detecting SNP epistatic interactions, such as CSE [29], MACOED [37], epiACO [25], and NHSA-DHSC [31], can perform only one task (detecting a single *k*th-*order* epistatic interaction) in each run and, therefore, must be run n times to perform n tasks (detecting k_1-*order*, k_2-*order*, …, k_n-*order* epistatic interactions). To collaboratively perform multiple detection tasks simultaneously, a multitasking HS algorithm (named MTHSA-DHEI) is developed for detecting high-*order* epistatic interactions in this study. The contributions of our work can be summarized as follows:

(1) A new multitask-based HS algorithm is proposed for detecting k_1-*order*, k_2-*order*, …, k_n-*order* SNP epistatic interactions simultaneously. The proposed algorithm is divided into three stages: searching, screening and verifying. The search stage aims to reduce the computational burden. The purpose of screening and verifying is to improve detection result accuracy.
(2) Unified coding is adopted to represent k_1-*order*, k_2-*order*, …, k_n-*order* combinations. For all tasks, the

solutions (SNP combinations) are encoded with the same length, which is equal to the number of SNPs in the highest-order SNP epistatic interaction, and this encoding scheme facilitates knowledge transfer between tasks. Knowledge transfer between tasks can significantly accelerate the detection of high-*order* SNP epistatic interactions from high-dimensional genome datasets.

(3) To improve the capability of identifying various models and discriminating *k*th-*order* SNP epistatic interactions from non-functional *k*th-*order* SNP combinations, four complementary association evaluation functions (Bayesian network, mutual entropy (ME), LR, and normalized distance with joint entropy (ND-JE)) are integrated as objective functions of the multitasking HS.

(4) **T** harmony memories (HM_1, HM_2, …, $HM_{\mathbf{T}}$) are employed to memorize the potential SNP epistatic combinations for **T** tasks, and four elite harmony sets (EHS_1, EHS_2, EHS_3, and EHS_4) are used for each task to record the elite solutions of four evaluation functions, with the aims of reducing the preference of a single evaluation function for a particular disease model and enhancing the global search ability.

The rest of this paper is arranged as follows. "Preliminary and related work" presents the related work and preliminaries. The proposed method is introduced in detail in "Proposed algorithm". The experiments performed on simulation datasets and real datasets are given in "Simulation experiments". The subsequent sections are the conclusion and discussion.

Preliminary and related work

Let $X = \{x_1, x_2, \cdots, x_N\}$ indicate N SNP markers for n individuals and $Y = \{y_1, y_2, \cdots, y_J\}$ denote disease status (J is the number of disease statuses). The homozygous major allele, heterozygous allele and homozygous minor allele in the sample dataset are defined as 0, 1 and 2, respectively. For a *k*th-*order SNP* combination, there are $I = 3^k$ genotype combinations. n_i is the number of samples in the dataset with SNP loci having the value of the *i*th genotype combination, and n_{ij} represents the number of samples with the *i*th genotype combination that are associated with disease state y_j.

Definition (*high-order SNP epistatic interaction*). Let $X_k = \{x_{s_1}, x_{s_2}, \cdots, x_{s_k}\}$ $(1 < k < N, x_{s_i} \in X)$ be a kth-*order* SNP combination. $f(X_k, Y)$ is a function for scoring the association between X_k and disease state Y. A kth-*order* SNP combination X_k is jointly associated with Y if and only if $\forall X' \subset X_k \wedge f(X_k, Y) \succ f(X', Y)$, where $\succ$ is defined for comparing the strength of association with the disease. X_k is said to be strongly associated with Y if $f(X_k, Y) > \theta$ (θ is the threshold value for determining the association with disease status). A *k*th-*order* SNP epistatic interaction occurs if and only if a *k*th-*order* SNP combination X_k is truly a disease-causing SNP combination associated with Y.

Multitasking optimization model

Multitasking optimization aims to solve K optimization problems simultaneously [41], and its goal is to concurrently optimize all K tasks. Let the K tasks be maximization problems. The optimization model can be expressed as follows:

$$\left\{X_1^*, X_2^*, ..., X_K^*\right\} = \left\{\arg\max f_1(X), .., \arg\max f_m(X)\right\}$$

where the objective function is defined as $f_i : S_i \to \mathbb{R}$ and $X_i^* \in S_i$ is the optimal solution of objective function f_i in the feasible space of S_i.

Evolutionary multitasking optimization (EMO) has received much attention in recent years in relation to implicit parallel population-based optimization algorithms to search multiple decision spaces of multiple optimization problems [42]. The evolutionary multitasking algorithm can significantly accelerate convergence for multiple complex optimizations by transferring learning between tasks. It has been applied in the fields of engineering and science computing. Li et al. employed a multifidelity evolutionary multitasking method to extract hyperspectral endmembers [43]. Feng et al. proposed evolutionary multitasking to solve the capacitated vehicle routing problem [44] consisting of a weighted learning process for capturing transfer mapping. Eneko Osaba et al. presented a novel adaptive metaheuristic algorithm to address evolutionary multitasking environments called the adaptive transfer-guided multifactorial cellular genetic algorithm (AT-MFCGA) [45]. Nguyen Thi Tam introduced evolutionary multitasking optimization to address the issues of relay node assignment for wireless single-hop sensor and multihop sensor networks in three-dimensional terrains [46]. To solve scheduling problems with batch distribution, Xu et al. presented multitasking optimization [47]. Gao et al. designed a transfer strategy based on the multidirectional prediction method to improve the performance of the multiobjective multitasking optimization approach [48]. Zhao et al. proposed a polynomial regression surface modelling approach based on multitasking optimization for rational basis function selection [49]. EMO can efficiently address multiple different optimization problems simultaneously, enhance the global search ability and improve the performance of each task via knowledge transfer between tasks [48].

In the detection of high-*order* SNP epistatic interactions, there may be an implicit relationship between kth-*order* SNP epistatic interactions and (kth + i)-*order* SNP epistatic interactions for the same disease.

For example, in a 5th-*order* SNP epistatic interaction $(x_1, x_2, x_3, x_4, x_5)$, the five single-SNP loci $x_i (i = 1, 2, ..., 5)$ and all 2nd-*order* SNP combinations $\left(x_i, x_j\right)(i \neq j)$ may have no explicit associations with disease status, while some 3rd-*order* SNP combinations, such as (x_1, x_2, x_4) and (x_2, x_3, x_5), show associations with disease status, which can guide the search algorithm to identify the 5th-*order* SNP epistatic interaction by transferring learning between the task of detecting 3rd-*order* epistatic interactions and that of detecting 5th-*order* epistatic interactions. In the task of detecting 3rd-*order* SNP epistatic interactions, some 3rd-*order* SNP combinations with very weak associations with the disease but no disease-causing SNP interactions may be part of the 5th-*order* epistatic interaction. Conversely, some 5th-*order* SNP combinations may contain functional loci for 3rd-*order* SNP interactions. Therefore, a multitasking optimization model is well suited for accelerating the detection of high-*order* SNP epistatic interactions through knowledge transfer between multiple tasks.

Multitasking optimization model for detecting high-*order* SNP epistatic interactions

The multitasking optimization model for detecting k_1 - order, k_2 - order, $\cdots$, k_m - order SNP epistatic interactions can be expressed as Eq. (1).

$$\begin{aligned}&\left\{X_{k_1}^*, \cdots, X_{k_m}^*\right\}\\&=\left\{\arg\max f_1(X, Y, k_1), \cdots, \arg\max f_m(X, Y, k_m)\right\},\end{aligned} \tag{1}$$

where $X_{k_i}^* (i = 1, 2,\ldots,m)$ indicates a k_i - order SNP epistatic interaction and $f(X, Y, k)$ denotes the objective function for evaluating the association between kth-*order* SNP combination X_k and disease status Y.

Discrimination functions for evaluating the associations between SNP combinations and disease status

Due to the small sample size and diversity of disease models, it is very difficult to discriminate kth-*order* SNP epistatic interactions from all kth-*order* SNP combinations on a genome-wide scale. Conventional evaluation methods (such as mutual information and Bayesian networks) cannot identify all disease models well. Almost all evaluation methods can correctly discriminate only a portion of disease models. In this study, four discrimination functions with low computational costs are employed to enhance the discrimination ability.

Bayesian-network-based K2 score. The Bayesian-network-based K2 score is a statistical method for describing relationships using a directed acyclic graph (DAG) $G = (V, E)$ [50]. It is a lightweight computing method and has high discrimination precision for evaluating the association between a kth-order SNP combination and disease status; it can be expressed as Eq. (2):

$$\text{K2 - Score}_{\log} = \sum_{i=1}^{I}\left(\sum_{k}^{n_i+1}\log(k) - \sum_{j=1}^{J}\sum_{s=1}^{n_{ij}}\log(s)\right). \tag{2}$$

The larger the K2 - $\text{Score}_{\log}$ value is, the greater the association between a SNP combination and disease status.

ME score. The ME score aims to calculate the contribution of a kth-order SNP combination X to disease status Y, defined as in Eq. (3) [51],

$$\begin{aligned}&\text{ME - score} = I(Y|x_1, ..., x_k)\\&=H(Y) + H(x_1, ..., x_k) - H(Y, x_1, ..., x_k),\end{aligned} \tag{3}$$

where $H(x)$(see Eq. (4)) denotes the Shannon entropy of x and $H(x_1, x_2, ..., x_k)$ represents the joint entropy of multiple variables $(x_1, x_2, ..., x_k)$.

$$H(Y) = -\sum_{y=0}^{J} p(y) \times \log_2(p(y)), \tag{4}$$

$$\begin{aligned}H(x_1, ..., x_k) = -\sum_{x_1=0}^{I}\cdots\sum_{x_k=0}^{I} p(x_1, ..., x_k)\\ \times \log_2 p(x_1, ..., x_k).\end{aligned} \tag{5}$$

LR score. The LR score is employed as a related measure to identify the likelihood difference between a kth-order SNP epistatic interaction and a kth-order SNP combination that is not involved in the disease process [52, 53] as shown in Eq. (6):

$$\begin{aligned}\text{LR - score} &= 2\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij} \ln\left(\frac{o_{ij}}{e_{ij}}\right)\\&= 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right),\end{aligned} \tag{6}$$

where o_{ij} and e_{ij} represent the observed number and expected number of phenotypes, respectively, when a phenotype takes the ith disease state and a SNP combination takes the jth genotype. The expected number e_{ij} can be obtained based on the Hardy–Weinberg principle [45, 58].

ND-JE score. The ND-JE score is defined as the normalized distance with joint entropy [32], which aims to uncover clues for detecting high-order epistatic models with very weak or no marginal effects, defined as in Eqs. (7)–(9):

$$\text{ND\_ JE - score} = ND(X)/JE(X_{\text{control}}), \tag{7}$$

$$ND(X) = \frac{\sum_{j=1}^{J}\sum_{i=0}^{I}\sqrt{\left(n_i^{j,\text{control}} - n_i^{j,\text{case}}\right)^2}}{\sum_{i=0}^{I}\left|n_i^{\text{control}} - n_i^{\text{case}}\right|}, \tag{8}$$

$$JE(X_{\text{control}}) = -\sum_{i=1}^{I}\frac{n_i^{\text{control}}}{n^{\text{control}}}\log\frac{n_i^{\text{control}}}{n^{\text{control}}}, \tag{9}$$

where $X = (x_1, x_2, \cdots, x_k)$ is a kth-*order* SNP combination for all samples (including case and control samples); X_{control} indicates a kth-*order* SNP combination for only control samples; $n_i^{j,\text{control}}$ and $n_i^{j,\text{case}}$ denote the numbers of control samples and case samples in the dataset, respectively, with the jth SNP locus taking the value of i (homozygous major allele 0, heterozygous allele 1 and homozygous minor allele 2); n_i^{control} and n_i^{case} represent the numbers of control samples and case samples in the dataset, respectively, with SNP combination X taking the value of the ith genotype combination; and n^{control} is the number of control samples.

The smaller the value of $ND(X)$ is, the larger the distribution difference (distance) between the case and control samples. The JE of the control samples is employed to normalize the distance. The main goal of ND-JE is to uncover a clue to guide the HS algorithm to detect potential disease-causing SNP combinations.

Harmony search algorithm

The HS algorithm mimics the process of new music improvisation by jazz musicians, who address unknown complex problems by exchanging information and learning between individuals in a group [54–56]. Musicians improvise their instruments' pitches to search for a perfect state of harmony. The HS algorithm is characterized by its simplicity, easy implementation, and powerful global search capabilities and has been widely applied in combination optimization problems on a large scale. (The standard HS algorithm is introduced in detail in Supplementary file 1.)

In HS, a candidate solution $X = (x_1, x_2, \ldots, x_{\mathbf{K}})$ is referred to as a harmony. A set of candidate solutions is referred to as a harmony memory (HM), which is similar to the memory of a tabu search (TS) algorithm and the population of a GA. The number of harmonies in an HM is called the harmony memory size (HMS). An HM is a matrix of *order* HMS $\times$ N or an augmented matrix of *order* HMS $\times$ $(N + 1)$ [50, 57] as in Eq. (10):

$$\begin{aligned}\text{HM} &= \left[\begin{array}{c|c} X^1 & f(X^1, Y, k) \\ X^2 & f(X^2, Y, k) \\ \vdots & \vdots \\ X^{\text{HMS}} & f(X^{\text{HMS}}, Y, k)\end{array}\right] \\ &= \left[\begin{array}{cccc|c} x_1^1 & x_2^1 & \cdots & x_{\mathbf{K}}^1 & f(X^1, Y, k) \\ x_1^2 & x_2^2 & \cdots & x_{\mathbf{K}}^2 & f(X^2, Y, k) \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ x_1^{\text{HMS}} & x_2^{\text{HMS}} & \cdots & x_{\mathbf{K}}^{\text{HMS}} & f(X^{\text{HMS}}, Y, k)\end{array}\right],\end{aligned} \tag{10}$$

where X^i $(i = 1, 2, \ldots, \text{HMS})$ is the i-th harmony in HM and $f(x^i)$ denotes the value of the objective function.

The worst harmony $X^{\text{id\_ worst}}$ in HM is iteratively updated by new harmony X^{new}, which is improvised through the following three operators:

(1) HM consideration performs a combination operation of HM with the probability harmony memory considering rate (HMCR).
(2) Pitching adjusts with probability PAR, which performs a local adjustment operation on X^{new}.
(3) Random consideration is performed with probability 1-HMCR, which introduces stochastic disturbance in a feasible search space to explore unknown space.

HS has been widely used to solve complex engineering and science optimization problems.

Proposed algorithm

Framework

Figure 1 shows the framework of our proposed MTHSA-DHEI algorithm.

MTHSA-DHEI is divided into three stages: the search stage, screening stage and verification stage. In the search stage, a multitasking HS is adopted to find potential SNP combinations that have a strong association with disease status. The G-test [30, 32, 59] statistical method is employed to test the significance level of the difference between control samples and case samples in the screening stage, and the SNP combinations with significance level p values larger than the threshold value θ_1 are discarded. In the verification stage, MDR [38] is further used to verify the classification ability.

The goal of this study is to develop a fast and effective search algorithm; therefore, the focus is on the design of the search stage.

In the search stage of MTHSA-DHEI, four scoring functions (Bayesian network-based K2 score, LR-based score, ME-based score and ND-JE-based score) are employed as

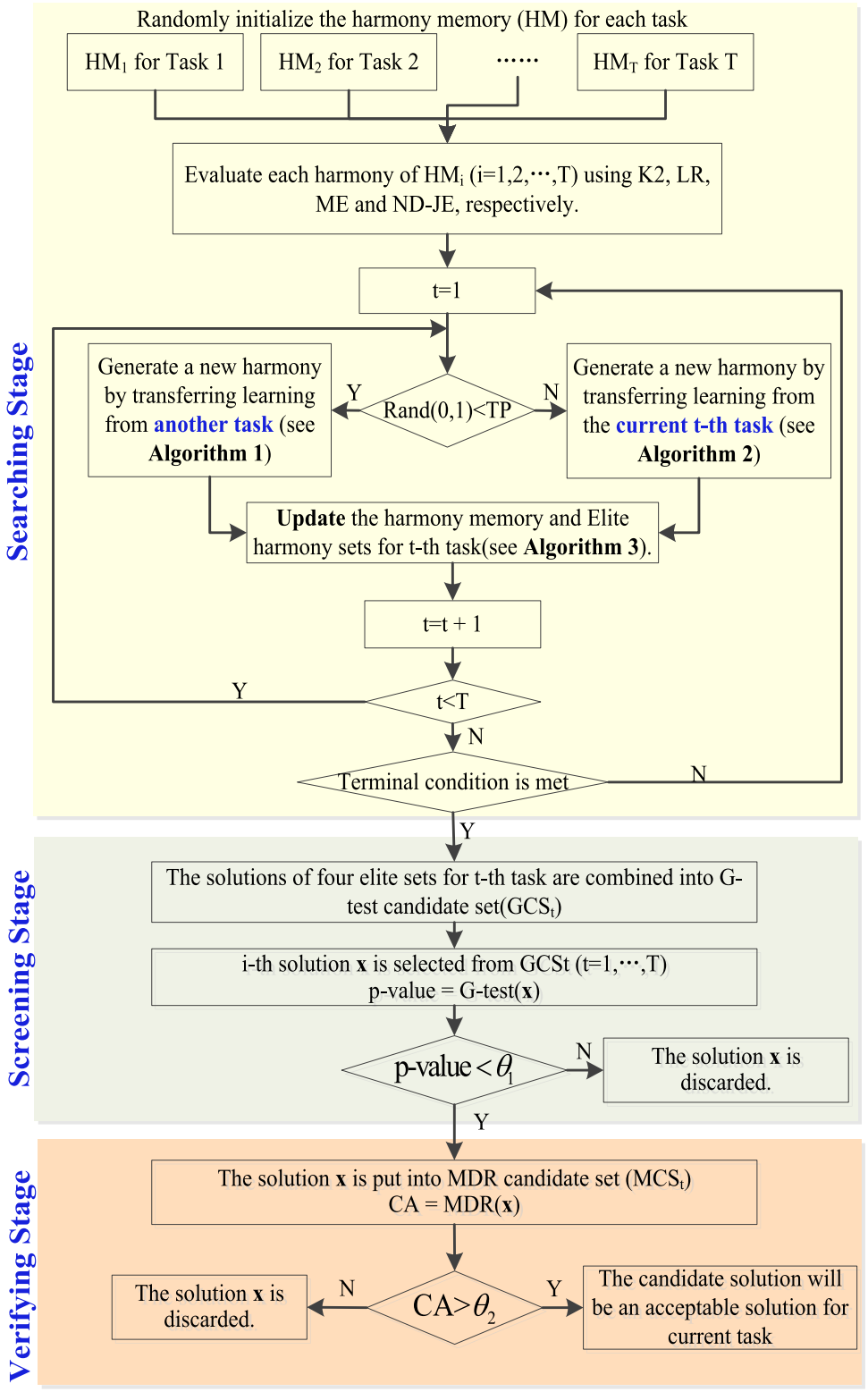

Fig. 1 The general flowchart of the proposed algorithm MTHSA-DHEI. In the search stage, the multitasking harmony search aims to discover some potential SNP combinations that have a strong association with disease status. In the screening stage, the G-test is employed to eliminate some SNP combinations without a significance level. MDR is used to verify the classification ability of candidate solutions in the 3rd stage

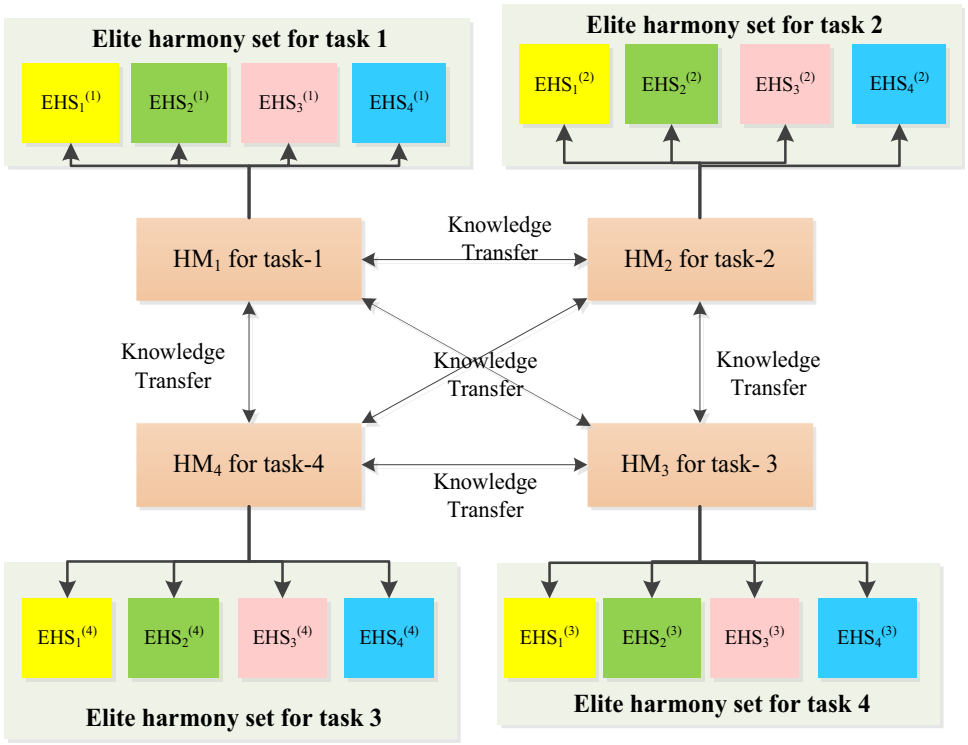

Fig. 2 Four tasks are separately employed to detect 2nd-*order*, 3rd-*order*, 4th-*order* and 5th-*order* SNP interactions simultaneously. The t-th task possesses an HM (HM_t) for recoding the potential SNP combinations and has four elite harmony sets ($EHS_1{}^t$, $EHS_2{}^t$, $EHS_3{}^t$ and $EHS_4{}^t$) for storing the elite solutions. In $EHS_1{}^t$, the association of SNP combinations is calculated with the K2 score based on a Bayesian network. The association of SNP combinations is calculated with the LR in $EHS_2{}^t$. In $EHS_3{}^t$ and $EHS_4{}^t$, the association of SNP combinations is separately calculated based on the ME and ND-JE scores. The tasks exchange information with each other to improve the search capability and speed

objective functions to improve the ability to discriminate SNP interactions with nonfunctional SNP combinations. T tasks are employed to detect 2nd-*order*, 3rd-*order*, …, Tth-*order*, (Tth + 1)-*order* SNP epistatic interactions simultaneously. As shown in Fig. 2, four tasks are concurrently employed to detect 2nd-*order*, 3rd-*order*, 4th-*order* and 5th-*order* SNP interactions, in which the t-th task has an HM and four elite harmony sets ($EHS_1{}^t$, $EHS_2{}^t$, $EHS_3{}^t$ and $EHS_4{}^t$). Each harmony in the HM has four association scores (K2 score, LR score, ME score and ND-JE score). Each harmony in the elite harmony set has only one association score. The K2 score, LR score, ME score and ND-JE score are separately adopted by $EHS_1{}^t$, $EHS_2{}^t$, $EHS_3{}^t$ and $EHS_4{}^t$. Unified coding is applied to the harmonies and elite harmony sets of all tasks, which is intended to allow the harmonies among the K tasks to transfer knowledge from each other and further improve detection speed.

In *MTHSA-DHEI*, all tasks employ the same code length, but only the previous $t + 1$ values are considered to be the solution for the t-th task. For example, task 1 aims to detect 2nd-*order* SNP interactions. Only the association of the 2nd-*order* SNP combination $\left(x_1^i, x_2^i\right)$ in $X^i (i = 1, 2, ..., HMS)$ is calculated between $\left(x_1^i, x_2^i\right)$ and disease status Y, and other SNPs are ignored, as follows:

$$\mathrm{HM}_1 = \left[\begin{array}{cccc|c} x_1^1 & x_2^1 & x_3^1 \cdots & x_{\mathbf{K}}^1 & f\left(\left(x_1^1, x_2^1\right), Y, 2\right) \\ x_1^2 & x_2^2 & x_3^2 \cdots & x_{\mathbf{K}}^2 & f\left(\left(x_1^2, x_2^2\right), Y, 2\right) \\ \vdots & \vdots & \vdots \cdots & \vdots & \vdots \\ x_1^{\mathrm{HMS}} & x_2^{\mathrm{HMS}} & x_3^{\mathrm{HMS}} \cdots & x_{\mathbf{K}}^{\mathrm{HMS}} & f\left(\left(x_1^{\mathrm{HMS}}, x_2^{\mathrm{HMS}}\right), Y, 2\right) \end{array}\right].$$

In Fig. 1, the general flow of the search stage is presented, in which Algorithm 1 and Algorithm 2 introduce the improvisation of new harmonies based on knowledge transfer from other tasks and the current task, respectively. Algorithm 3 presents the process of updating the harmony memory and elite harmony sets ($EHS_1{}^t$, $EHS_2{}^t$, $EHS_3{}^t$ and $EHS_4{}^t$) of the t-task.

Improvising new harmonies with knowledge transfer

In the proposed MTHSA-DHEI approach, improvising a new harmony for the current t-th task has two steps: (1) knowledge transfer from other tasks and (2) generation of a new solution in the current task by using HM_t, $EHS_1{}^t$, $EHS_2{}^t$, $EHS_3{}^t$ and $EHS_4{}^t$. Figure S1 shows an example in which task 2 transfers knowledge to task 1 (see Supplementary file 1).

In the 1st method, the new harmony is improvised using three classical operators of HS, but the knowledge is from another task $t_r \in \{1, 2, \cdots, T\}$, $t_r \neq t$, which aims to obtain

the information from another task t_r. If task t_r has a higher order than the current task, it may carry one or more functional SNPs that were missing in the current task. Conversely, if task t_r has a lower order than the current task, some clues (SNPs with marginal effects) that can help the current task accelerate the detection of epistatic SNP interactions may be found.

Algorithm 1 describes the steps of improvising a new harmony $X^{\text{new}} = \left(x_1^{\text{new}}, x_2^{\text{new}}, \ldots, x_K^{\text{new}}\right)$ for the t-th task by transferring learning from the t_r-th task ($t \neq t_r$), in which the HM consideration is from the four EHSs of the t_r-th task; there are three strategies with equal probability of pitch adjustment X^{new}. For X^{new}, if $x_i^{\text{new}} = x_j^{\text{new}} (i \neq j)$, the value of x_i^{new} or the value of x_j^{new} will be randomly regenerated from the search space. X^{new} will be randomly regenerated if it exists in the HM or in $\text{EHS}_r^t (r = 1, 2, 3, 4)$.

Algorithm 1. Improvise a new harmony $X^{new} = \left(x_1^{new}, x_2^{new} \cdots, x_K^{new}\right)$ for the t-th task by knowledge transfer from t_r-task.

1: Randomly select one task $t_r \in \{1,2,\cdots,\mathrm{T}\}, t_r \neq t$.
2: Randomly generate a random integer $R \in \{1,2,3,4\}$.
3: **For** i = 1 to K
4: **If** rand(0,1) < HMCR
5: $x_i^{new} \leftarrow x_a^{\mathrm{EHS}_{\mathrm{R}}^{t_r,b}} \in \mathrm{EHS}_R^{t_r}$
6: **If** rand(0,1) < PAR
7: r ← randInt({1,2,3}).
8: **If** r==1
9: $x_i^{new} \leftarrow x_{a_1}^{\mathrm{EHS}_{\mathrm{R}}^{t_r,best}} \in \mathrm{EHS}_{\mathrm{R}}^{t_r}$
10: **Else if** r==2
11: $L = x_{a_1}^{\mathrm{EHS}_{\mathrm{R}_0}^{t_r,best}} - \mathrm{HM}_{t_r}(b_2, a_1)$
12: $x_i^{new} \leftarrow x_i^{new} + F \times rand(0,1) \times L$
13: **Else**
14: $L = x_{a_1}^{\mathrm{EHS}_{\mathrm{R}}^{t_r,h_1}} - \mathrm{HM}_{t_r}(b_2, a_1)$
15: $x_i^{new} \leftarrow x_i^{new} + F \times rand(0,1) \times L$
16: **End**
17: **EndIf**
18: **Else**
19: $x_i^{new} = r \in \{1,2,\cdots,N\}$
20: **EndIf**
21: **EndFor**
22: **While** $X^{new} \in \mathrm{HM}_t \,||\, X^{new} \in \mathrm{EHS}_r^t$
23: $J = i \in \{1,2,\cdots,k\}$
24: $x_j^{new} = r \in \{1,2,\cdots,N\}$
25: **EndWhile**

In the 2nd method, a new harmony is improvised through the components (HM_t, $\text{EHS}_1{}^t$, $\text{EHS}_2{}^t$, $\text{EHS}_3{}^t$, and $\text{EHS}_4{}^t$) of the current t-th task. HM_t is for harmony memory consideration with probability HMCR. The best harmonies of four elite harmony sets ($\text{EHS}_1{}^t$, $\text{EHS}_2{}^t$, $\text{EHS}_3{}^t$, and $\text{EHS}_4{}^t$) are the focus of consideration when employing the pitch adjustment operator.

Algorithm 2. Improvise a new harmony $X^{new} = \left(x_1^{new}, x_2^{new} \cdots, x_{\mathbf{K}}^{new}\right)$ through the components (HM_t, $\text{EHS}_1{}^t$, $\text{EHS}_2{}^t$, $\text{EHS}_3{}^t$, $\text{EHS}_4{}^t$) of the t-th task

1: **For** i = 1 to K
2: **If** rand(0,1) < HMCR
3: $x_i^{new} \leftarrow \mathrm{HM}_t(b, a)$
4: **If** rand(0,1) < PAR
5: r ← generate a random integer in {1,2,3}
6: **If** r==1, then $x_i^{new} \leftarrow x_{a_1}^{\mathrm{EHS}_{\mathrm{R}}^{t,best}} \in \mathrm{EHS}_R^t$
7: **Else if** r==2
8: $L = x_{a_1}^{\mathrm{EHS}_{\mathrm{R}}^{h_1}} - \mathrm{HM}_t(b_1, a_1)$
9: $x_i^{new} \leftarrow x_i^{new} + F \times rand(0,1) \times L$
10: **Else**
11: $L = x_{a_1}^{\mathrm{EHS}_R^{t,best}} - \mathrm{HM}_t(b_2, a_1)$
12: $x_i^{new} \leftarrow x_i^{new} + F \times rand(0,1) \times L$
13: **EndIf**
14: If $x_i^{new} > \mathrm{N} \,||\, x_i^{new} < 1$
15: $x_i^{new} \leftarrow r \in \{1,2,\cdots,N\}$
16: **End**
17: **End**
18: **Else**
19: $x_i^{new} = r \in \{1,2,\cdots,N\}$
21: **EndIf**
22: **EndFor**
22: **While** $X^{new} \in \mathrm{HM}_t \,||\, X^{new} \in \mathrm{EHS}_r^t \quad (r = 1,2,...,4)$
23: $J = i \in \{1,2,\cdots,k\}$
24: $x_j^{new} = r \in \{1,2,\cdots,N\}$
25: **EndWhile**

Algorithm 2 describes the steps of improvising a new harmony $X^{\text{new}} = \left(x_1^{\text{new}}, x_2^{\text{new}}, \ldots, x_K^{\text{new}}\right)$ for the t-th task by self-learning from HM_t and $\text{EHS}_r^t (r = 1, 2, 3, 4)$.

In Algorithm 1 and Algorithm 2, $a \in \{1, \cdots, K\}$, $a_1 \in \{1, \cdots, K\}$, $b \in \{1, \cdots, \text{HMS}\}$, $b_1 \in \{1, 2, \cdots, \text{HMS}\}$, $b_2 \in \{1, \ldots, \text{HMS}\}$, $R \in \{1, 2, 3, 4\}$, $R_0 \in \{1, 2, 3, 4\}$ and $r \in \{1, 2, 3, 4\}$ are all randomly generated integers, and $\text{HM}_{t_r}(i, j)$ denotes the j-th note (variable) of the i-th harmony in HM_{t_r}. EHS_r^t is the r-th elite harmony set of the t-th task. $x_a^{\text{EHS}_r^t, b}$ denotes the a-th SNP value of the b-th harmony in the r-th elite harmony set of the t-th task. F is the scale factor for adjusting the step of the local search.

In Algorithm 1 and Algorithm 2, the scale factor F is important for the performance of the proposed MTHSA-DHEI method. It is analysed in the simulation experiment described in "Simulation experiments".

Update the harmony memory and elite harmony sets

For each new harmony generated for the t-th task, HM_t, $EHS_1{}^t$, $EHS_2{}^t$, $EHS_3{}^t$ and $EHS_4{}^t$ are considered to be updated. The update rate of HM_t is associated with FEs. Algorithm 3 describes the update operator in detail.

Algorithm 3: Update HM_t, EHS_k^t **(k=1, 2, 3, 4) for the t-th task**

1: **(1) Calculate the association between** X^{new} **and** Y.
2: $Score^{new} = \left[sc^{new,1}, sc^{new,2}, sc^{new,3}, sc^{new,4}\right] = f(X^{new}, Y, D^t)$
3: $sc^{new,i} = sc^{new,i} / \max\_score^i \quad (i = 1,2,3,4)$
4: **(2) Update the** HM_t.
5: **For i=1 to HMS**
6: $C = sum\left(Score^{new} > Score^i_{HM_t}\right)$
7: **If** $C>2 \,||\, sc^{new,4} > sc^{i,4}_{HM_t}$ && $r < (1 - FEs/maxFEs)$
8: $HM_t(i,:) \leftarrow X^{new}$; break;
9: **EndIf**
10: **EndFor**
11: **(3) Update elite harmony sets.**
12: **For k = 1 to 4**
13: **If** size (EHS_k) < EliteSize
14: put X^{new} into EHS_k.
15: **Elseif** $sc^{new,k} > sc^{id\_worst}_{EHS_k^t}$
16: $X^{EHS_k^t, id\_worst} \leftarrow X^{new}$
17: **EndIf**
18: **EndFor**

C is the number of scores in $Score^{new}$ that are higher than the corresponding scores in $Score^i$.
FEs are the number of SNP combinations that have been evaluated to date for their association with disease status.
$X^{EHS_k^t, id\_worst}$ is the worst individual in EHS_k^t.

In Algorithm 3, the i-th fitness value $sc^{new,i}$ of X^{new} is divided by max_ score^i to normalize the fitness value to the interval [1]. The value of max_ score^i is the maximum value of the i-th scoring function in the initial population, and its value is not changed during iterations. The condition $C > 2 \,||\, sc^{new,4} > sc^{i,4}_{HM_t}$ &&rand(0, 1) < (1 − FEs/maxFEs) is critical. $C > 2$ means that the i-th harmony $HM_t(i, :)$ of HM_t is replaced by X^{new} only when at least two scores of X^{new} are higher than the corresponding scores of $HM_t(i, :)$.$sc^{new,4} > sc^{i,4}_{HM_t}$ && rand(0, 1) < (1 − FEs/maxFEs) indicates that $HM_t(i, :)$ can be replaced by X^{new} with probability (1 − FEs/maxFEs) if $sc^{new,4} > sc^{i,4}_{HM_t}$, which means that the ND-JE score is different from the other three scores. In this work, the goal of the ND-JE score is to discover some clues to guide the algorithm to locate SNP interactions with no marginal effect on disease status.

In the proposed MTHSA-DHEI algorithm, four complementary discriminating functions (evaluation functions) are adopted to calculate the associations between high-order SNP combinations and disease status, with the aim of improving the ability to identify various diseases, and have the following benefits:

(1) All four evaluation methods are lightweight. For a kth-order SNP combination, the four scoring values can be calculated simultaneously by counting only the values of n_i and n_{ij} (i = 1, 2,…,I; j = 1,2), and the calculations are not additive.
(2) The four evaluation methods are complementary to each other. The K2 score has high power for detecting SNP interactions and is superior in discriminating certain disease models with weak marginal effects. However, it has low accuracy for interaction models with low minor allele frequencies (MAFs) and low genetic heritability (H^2). The ME score aims to calculate the contribution of a kth-*order* SNP combination to disease status. The LR score aims to discover the likelihood difference between a functional SNP combination and a nonfunctional SNP combination via statistical theory, and it has good adaptability to unknown disease models. The ND_JE score aims to guide the HS to uncover clues for detecting high-*order* epistatic interactions.

Simulation experiments

To investigate the performance of the proposed MTHSA-DHEI method, four 4th-*order* and eight 5th-*order* epistatic interaction models with marginal effects (EIMEs), eight high-*order* epistatic interaction models with no marginal effects (EINMEs) (including two 3rd-*order* models, three 4th-*order* models and three 5th-*order* models) and one real disease dataset (age-related macular degeneration, AMD) were tested. The experimental results were compared with the results of four high-*order* epistatic interaction detection algorithms: CSE, epiACO, NHSA-DHSC and MP-HS-DHSI. All experiments were performed on a Windows 10 64-bit system with an Intel(R) Core (TM) i7-8700 CPU @3.2 GHz, 16 GB memory, and all the program codes were written and run in MATLAB R2018a.

Evaluation criteria for performance

(1) Power is a measure of the capability of identifying a *k*th-*order* SNP epistatic interaction from genomic data and is expressed as

$$\text{Power} = \frac{\#S}{\#T},$$

where $\#S$ is the number of true *k*th-*order* epistatic interactions found from the simulation datasets, in which a total of $\#T$ true *k*th-*order* epistatic interactions are available.

Note that in this work, power is used mainly to evaluate the search ability of the proposed method. If the disease-causing SNP combination (epistatic interaction) has been found within the specified iterations (maximum number of objective functions evaluated, *maxFEs*), the search is considered successful.

(2) FEs represent the number of evaluations of the associations between *k*th-*order* SNP combinations and disease status until the kth-*order* SNP epistatic interaction is found or the terminal condition of the algorithm is met. In simulation experiments, the search is stopped immediately when the *k*th-*order* SNP epistatic interaction is found, and FEs is the number of SNP combinations that have been evaluated for their association with disease status. In this work, the aim of the FEs is to measure the capability of the algorithm to reduce the computational burden.

(3) Runtime denotes the mean runtime that an algorithm takes to detect the *k*th-*order* SNP epistatic interaction before the algorithm is terminated, and it is intended to measure the time cost of detecting high-*order* SNP epistatic interactions.

To further investigate the reliability of the results obtained in the search stage, the G-test method is adopted to screen out the SNP combinations that differ significantly between case and control samples $\left(P\text{value} < \max\left(10^{-8}, 0.05/C_N^k\right)\right)$, and MDR is then used to verify the classification accuracy of the SNP combinations selected by the G-test [59]. The false discovery rate (FDR) and F1-score are employed to further evaluate the reliability of the results.

(4) The FDR is defined as:

$$\frac{FP}{FP + TP},$$

where FP and TP represent the false-positive rate and true-positive rate, respectively.

(5) The *F*1-score can be expressed as follows:

$$\text{recall} = \frac{TP}{TP + FN},$$

$$\text{precision} = \frac{TP}{TP + FP},$$

$$F1 - \text{score} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}.$$

Datasets

(1) EINME datasets. Eight EINMEs were employed to test the capability of detecting high-*order* epistatic interactions with no marginal effect. For each EINME, 1500 control samples and 1500 case samples for the functional SNPs were generated using a multiobjective optimization algorithm that aims to maximize the joint effects of *k*-SNPs, minimize the marginal effects of individual SNPs and limit Hardy–Weinberg equilibrium (HWE) constraints [60]. The samples of nonfunctional SNPs were randomly generated according to HWE. To investigate the performance of the proposed algorithm, simulation datasets with 100 SNPs, 1 k SNPs and 10 k SNPs were generated for each EINME. The eight EINMEs are described in Table S1 of Supplementary file 1.

(2) EIME datasets. Four 5th-*order* additive EIMEs, four 5th-*order* threshold EIMEs and four 4th-*order* multiplicative EIMEs [61] were employed to test the performance of detecting epistatic interactions with marginal effects. For each model, 100 datasets with 2000 control samples and 2000 case samples that had 100 SNPs, 1000 SNPs and 10 k SNPs were separately generated using GAMETES software [62]. The parameters of the 12 EIMEs are *listed in* Table S2 of Supplementary file 1.

(3) AMD dataset. The AMD dataset contains 103,611 SNPs genotyped for 50 controls and 96 cases [63]. This experiment aims to detect 2nd-*order*, 3rd-*order*, 4th-*order* and 5th-*order* SNP epistatic interactions from the 103,611 SNPs for AMD. We conducted two experiments for AMD:

A. All 103,611 SNPs were detected to identify epistatic interactions.
B. Three widely reported SNPs (rs380390, rs1329428, and rs1363688) were first removed, and the rest of the SNPs were detected to identify epistatic interactions in which each SNP had a small effect on disease status.

Parameter analysis and settings

(1) Effect of parameters on the performance of MTHSA-DHEI

In this section, the effect of two important parameters (TP and F) on the performance of MTHSA-DHEI are investigated.

As shown in Fig. 3, for the EINMEs, when TP > 0.5, the power begins to drop gradually, and the FEs and runtime increase with an increasing TP value, except for EINME-4. However, as shown in Fig. 4, for the EIMEs, the FEs and

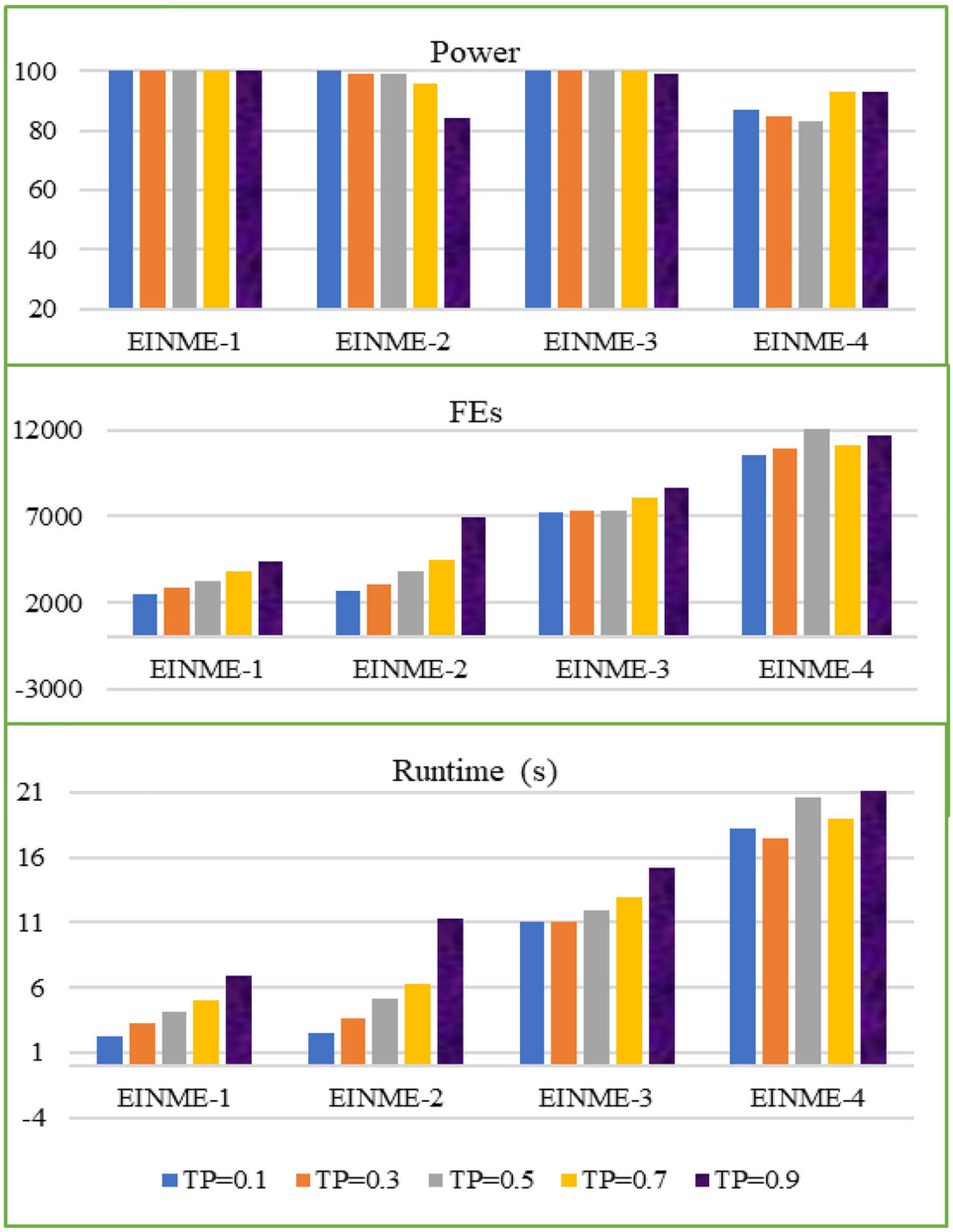

Fig. 3 Effect of TP on the performance of MTHSA-DHEI for detecting EINME interactions

runtime decrease with an increasing TP value, but the power remains constant. This result demonstrates that a larger TP value will decrease the performance of MTHSA-DHEI for detecting EINME interactions but enhance the performance for the detection of EIME interactions. With a compromise, we believe that TP = 0.5 is a better choice when we have unknown disease models.

Next, we investigate the effect of parameter F on the performance of the proposed method. Figures 5, 6 and 7 show the power, FEs and runtime of MTHSA-DHEI for values of parameter F from 2 to 20 (step = 2). MTHSA-DHEI has the same power values for the three EIMEs for all F values, and it has high power for EINMEs when F equals 10. As shown in Fig. 6 and Fig. 7, MTHSA-DHEI with a small F value ($F < 12$) requires more FEs and runtime to find epistatic interactions than MTHSA-DHEI with a large F value for EINMEs, but for the three EIMEs, the opposite result occurs. Therefore, we recommend that F be set to 10.

In addition, a θ_2 value set to 0.6 has the highest accuracy for all EINMEs and EIMEs. When $\theta_2 < 0.55$, the false-positive rate starts to increase, and when $\theta_2 > 0.65$, the false negative rate starts to increase. For PAR, the algorithm has a greater search speed and improved detection power when its value is in the interval [0.4, 0.7].

(2) Parameter settings

Fig. 4 Effect of TP on the performance of MTHSA-DHEI for detecting EIME interactions

The parameters of the algorithms are described in Table 1.

Experimental results and analysis

(1) EINME

Figure 8 shows the power, FEs and runtime used to detect eight high-*order* EINMEs using five intelligent search algorithms, and the results show that the power of MTHSA-DHEI exceeds that of the other four algorithms in all EINMEs except for MP-HS-DHSI, which has the same power value as MTHSA-DHEI on EINME-3, EINME-4 and EINME-6. Both MTHSA-DHEI and MP-HS-DHSI have much higher power than the other algorithms (Fig. 8a). As shown in Fig. 8b, MTHSA-DHEI took the fewest FEs among all five algorithms to find the *k*th-*order* SNP epistatic interactions except for EINME-3, EINME-4 and EINME-6. Except for EINME-5, EINME-7 and EINME-8, MTHSA-DHEI has a more than 99% success rate in detecting *k*th-*order* (k = 3,4,5) epistatic interactions from 100 SNPs with no more than 10,000 FEs, which is much lower than the number of FEs (C^3_{100}=161,700, C^4_{100}=3,921,225, C^5_{100}=75,287,520) obtained by an exhaustive search.

As shown in Fig. 8c, MP-HS-DHSI took the least time among the five algorithms on EINME-1, EINME-3, EINME-4 and EINME-6 to find the high-*order* SNP epistatic interactions. The runtime taken by the proposed MTHSA-DHEI method is slightly more than that of MP-HS-DHSI, but it is less than the runtime required by the other three algorithms. Importantly, in the simulation experiments, MP-HS-DHSI, NHSA-DHSC, epiACO and CSE were performed only on a *k*th-*order* epistatic interaction task (where *k* is the number of

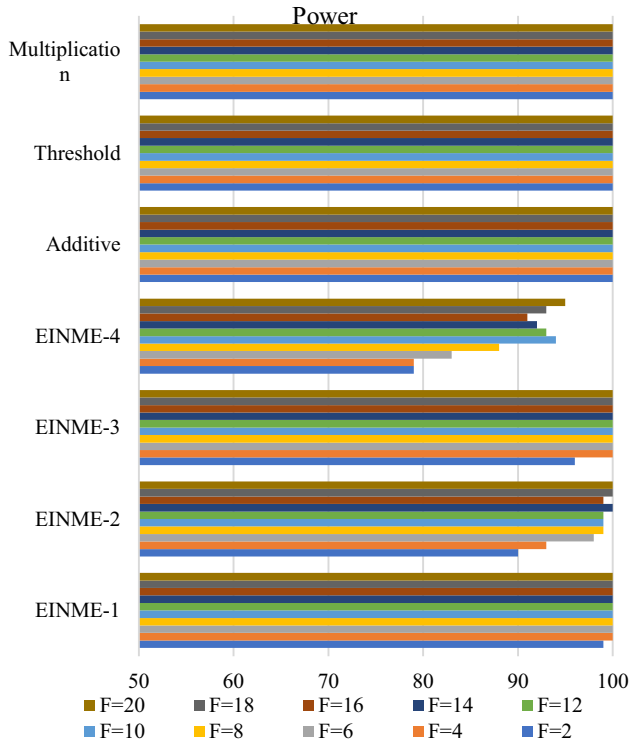

Fig. 5 Effect of parameter F on the power of MTHSA-DHEI

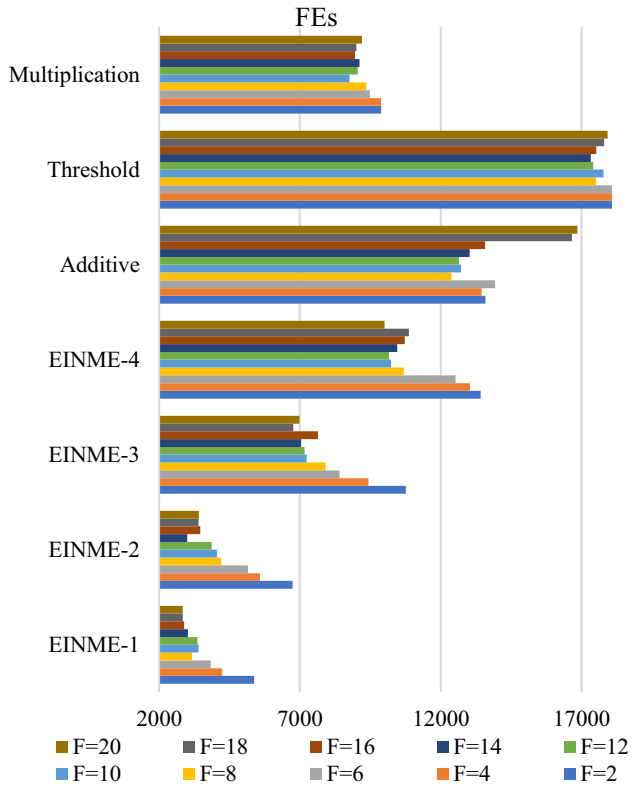

Fig. 6 Effect of parameter F on the FEs of MTHSA-DHEI

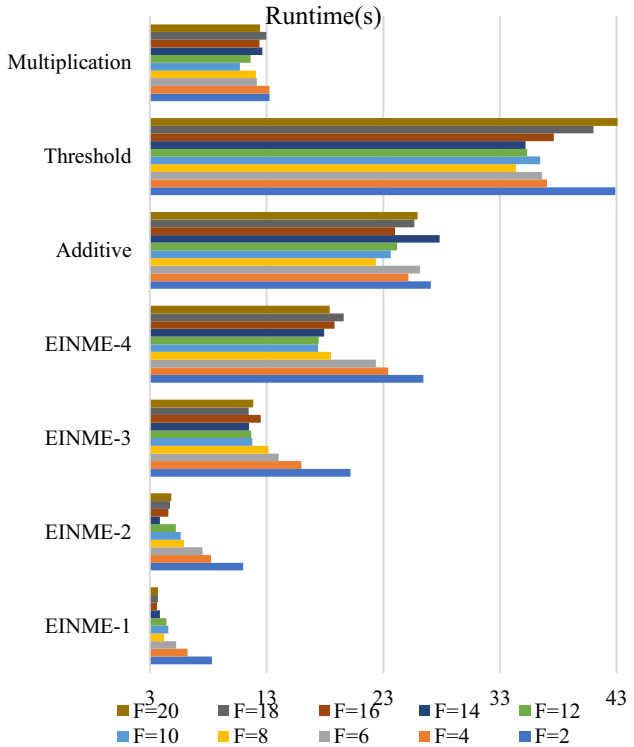

Fig. 7 Effect of parameter F on the runtime of MTHSA-DHEI

functional SNPs in an epistatic interaction); however, the proposed MTHSA-DHEI method aims to simultaneously detect 2nd-*order*, …, *k*th-*order* epistatic interactions, which consumes much of the computational cost of MTHSA-DHEI to detect potential 2nd-*order*, …, $(k-1)$-*order* epistatic interactions. Overall, MTHSA-DHEI has evident advantages over the other four approaches in eight EINMEs with 100 SNPs.

To further investigate performance as the number of SNPs increases, we conducted the proposed method on EINME datasets with 1 k and 10 k SNPs. The results are summarized in Table S3 (see Supplementary file 1). Figure 9 displays the change curves of the power, FEs, runtime and F1-score of the MTHSA-DHEI with an increasing number of SNPs, from which we can see that the power and F1-scores decrease rapidly and the FEs and runtime increase significantly when conducting EINME-5, EINME-7 and EINME-8; however, for the other five models, the changes in these metrics are not very significant. We found that in EINME-5, EINME-7 and EINME-8, the marginal effect of each functional SNP was very small, especially for EINME-8, and the joint effects could be seen only when three or more of the five functional SNPs were combined, making it very difficult to search for epistatic interactions among over thousands of SNPs. Compared with the exponential growth in the number of SNP combinations, the increases in FEs and runtime and the decrease in power are very small and acceptable.

(2) EIME

Table 1 Parameters of the five algorithms for detecting high-*order* epistatic interactions

Algorithm	Parameters
MTHSA-DHEI	HMCR = 0.98, PAR = 0.5, TP = 0.5, $T = K - 1$, $F = 10$ (1) For EINME and EIME simulation datasets, the HMS of each task is set to max(50, $K \times \min(N/10,200)$) (K is the highest-*order* SNP interactions that will be detected). MaxFEs = min(50 × K × N, 2 × K × 50,000), $\theta_1 = \max\left(10^{-8}, 0.05/C_N^k\right)$, $\theta_2 = 60\%$ (2) For the AMD dataset, HMS = 300, K = 4, MaxFEs = 5×10^7. $\theta_1 = 1 \times 10^{-7}$, $\theta_2 = 70\%$
MP-HS-DHSI	MaxFEs = 2 × K × 50,000, HMCR = 0.98, PAR = 0.35, HMS = 100, size of candidate sets is equal to 5
NHSA-DHSC	MaxFEs = 2 × K × 50,000, HMCR = 0.95, PAR = 0.35, HMS = 50, size of candidate sets is equal to 10. FEs in the local search is set to MaxFEs/4
epiACO	MaxFEs = 2 × K × 50,000, AntNumber = 200, evaporation coefficient $\rho = 0.2$; others are the same as the parameters of the author's source code
CSE	MaxFEs = 2 × K × 50,000, NestNumber = 100, SNP number in each group is set to 5; others are the same as the parameters of the author's source code

Table 2 summarizes the results (power, FEs, and runtime) of the five approaches in twelve high-*order* EIMEs, from which it can be clearly seen that the proposed MTHSA-DHEI method outperforms or is equivalent to the other four methods in terms of power. MTHSA-DHEI took fewer FEs than CSE, NHSA-DHSC and epiACO for almost all 12 models, and it had 100% power to detect epistatic interactions with very few FEs. Compared with MP-HS-DHSI, MTHSA-DHEI took fewer FEs on four multiplicative datasets (EIME-9-EIME-12) with 1000 SNP loci. MTHSA-DHEI took less runtime than CSE, NHSA-DHSC and epiACO, but it took more time than MP-HS-DHSI. However, the runtimes and FEs of MTHSA-DHEI are composed of the runtimes required to perform multiple tasks (detection of 2nd-*order*, 3rd-*order*, …, and *k*th-*order* SNP epistatic interactions), while the runtimes of the other four methods correspond to the time of performing only a single task (detection of kth-*order* SNP epistatic interactions). In summary, compared with the four excellent SISAs, MTHSA-DHEI has significant advantages in power, especially for more complex disease models (i.e., multiplicative models).

Experiments on the 12 EIME datasets with 100, 1 k and 10 k SNPs are conducted, and the results are summarized in Table S4 (see Supplementary file 1), which demonstrates that the proposed algorithm can maintain high power (1st and 2nd power) for all datasets with different SNPs. Moreover, its runtime and FEs increases are not very significant, where the 1st power denotes the ability to find functional epistatic interactions by the multitasking HS algorithm and the 2nd power is the number of epistatic interactions that pass the threshold value θ_1. However, for EIME-5, EIME-8, EIME-9 and EIME-10, the 3rd power is equal to zero because the ability to classify functional SNP epistatic interactions cannot pass the threshold value θ_2 (= 60%). When MDR was used to evaluate the four disease models, their average classification accuracy was equal to 56.8%, 58.6%, 56.3% and 58.4%. Comparing the EIMEs with EINMEs, each functional SNP in the EIMEs has an obvious marginal effect on disease status, which allows the HS algorithm to quickly locate the functional SNPs.

(3) AMD. The proposed MTHSA-DHEI method is adopted to simultaneously detect 2nd-*order*, 3rd-*order*, 4th-*order* and 5th-*order* epistatic interactions from AMD data, with 146 samples and 103,611 SNPs. A total of 526 2nd-*order* SNP combinations, 1059 3rd-*order* SNP combinations, 638 4th-*order* SNP combinations and 322 5th-*order* SNP combinations were found to be associated with AMD, of which 168 2nd-*order* SNP combinations (CA > 75%, p value < 1×10^{-7}), 631 3rd-*order* SNP combinations (CA > 80%, p value < 1×10^{-10}), 546 4th-*order* SNP combinations (CA > 85%, p value < 1×10^{-10}), and 285 5th-*order* SNP combinations met the significance level for the G-test [30, 32, 59], and the classification accuracy with MDR for each SNP combination was greater than 75%, 80%, 85% and 90% for the 2nd-order, 3rd-order, 4th-order and 5th-order combinations, respectively (see Supplementary file 2 for details). To better analyse the interactions among the identified SNPs, we employed Cytoscape software (https://cytoscape.org/) [67] to generate the interaction networks (see Fig. 10a, Figs. S2(a), S3(a) and S4(a)).

To detect epistatic interactions in which each SNP has a small effect on disease status, we removed three important SNPs (rs380390, rs1329428 and rs1363688) that have been widely reported to be associated with AMD, and MTHSA-DHEI was applied to the remaining SNPs. Twenty-four 2nd-order SNP combinations (CA > 75%, p value < 1×10^{-7}), 33 3rd-order SNP combinations (CA > 80%, p value < 1×10^{-10}), 56 4th-order SNP combinations (CA > 85%, p value < 1×10^{-10}) and 89 5th-order SNP combinations (CA > 88%, p value < 1×10^{-12}) were found to be strongly associated with AMD. Figure 10b, Figs. S2(b) and S3(b) show the interaction networks.

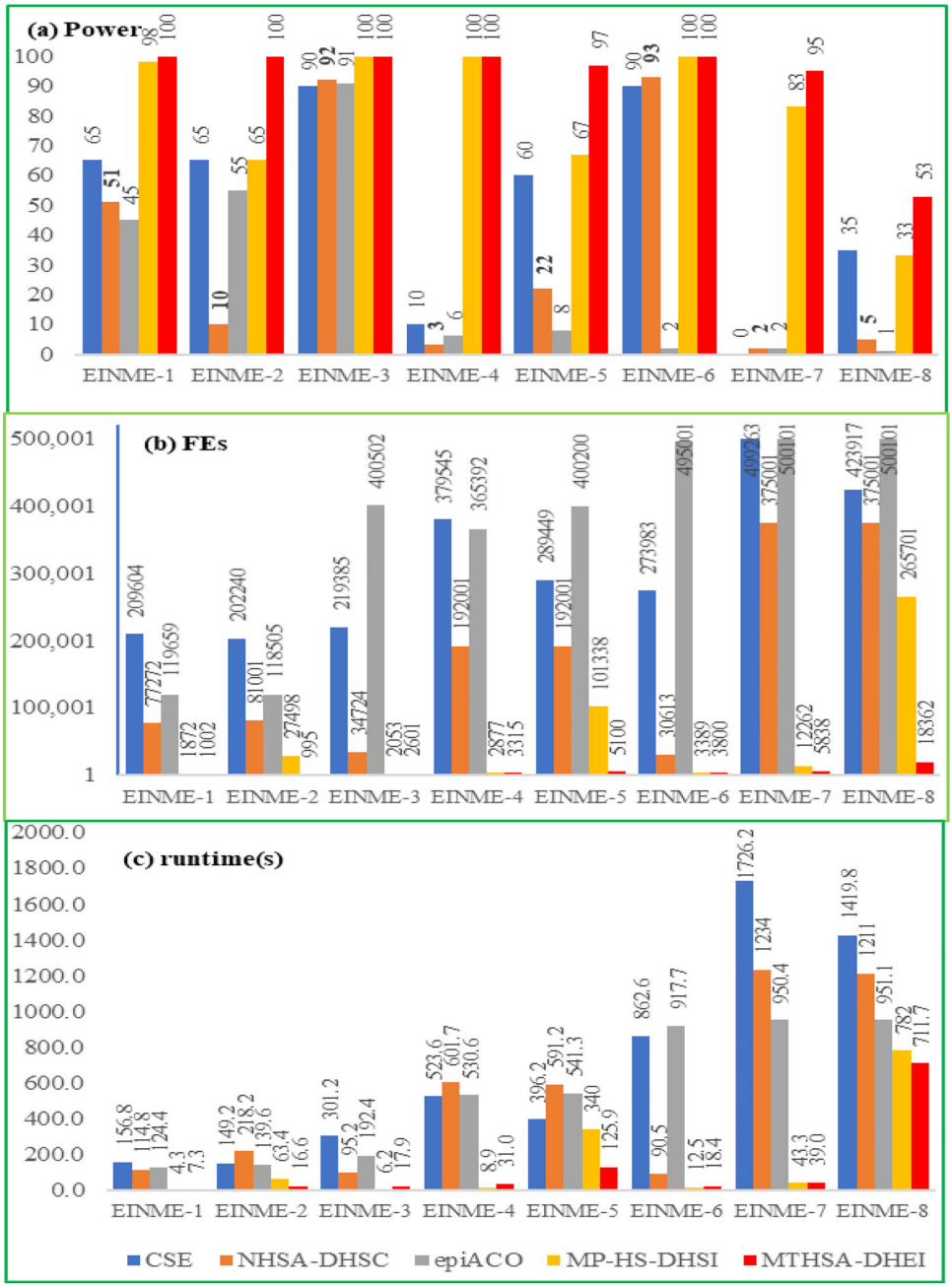

Fig. 8 Power, FEs and runtime of the five algorithms for detecting 8 EINMEs with 100 SNPs

Figure 10a shows the 2nd-*order* SNP combinations (p value $< 1 \times 10^{-7}$) with a classification accuracy greater than 75%, with SNPs rs380390, rs1329428 and rs10272438 shown to interact with many other SNPs. Both rs380390 and rs1329428 are in the *CFH* gene, which has been widely reported to be associated with AMD [3, 24, 26, 34, 63, 64]. rs10272438 is an intron variant of the *BBS9* gene that is associated with Bardet Biedl syndrome [65] and was reported in our previous study [31]. In Fig. 10b, rs10272438 is the only central node that interacts with 21 SNPs.

Figure S2(a) displays the interaction network of 3rd-*order* SNP combinations with a classification accuracy greater than 80%, of which the degrees of five central nodes, namely, rs380390, rs1363688, rs1329428, rs618499 and rs555174, are equal to 193, 124, 3, 47 and 13, respectively. In Figure S2(a), the SNPs rs380390 and rs1329428 are indicated to have important roles in 3rd-*order* SNP combinations, and rs1363688 (at position 174,609,731 of chromosome 15, not in a gene-coding region) [14, 31, 32] and rs618499 (in gene *ATM*) were reported to be associated with osteosarcoma [66] and AMD [14]. rs555174 (not in a gene-coding region) has

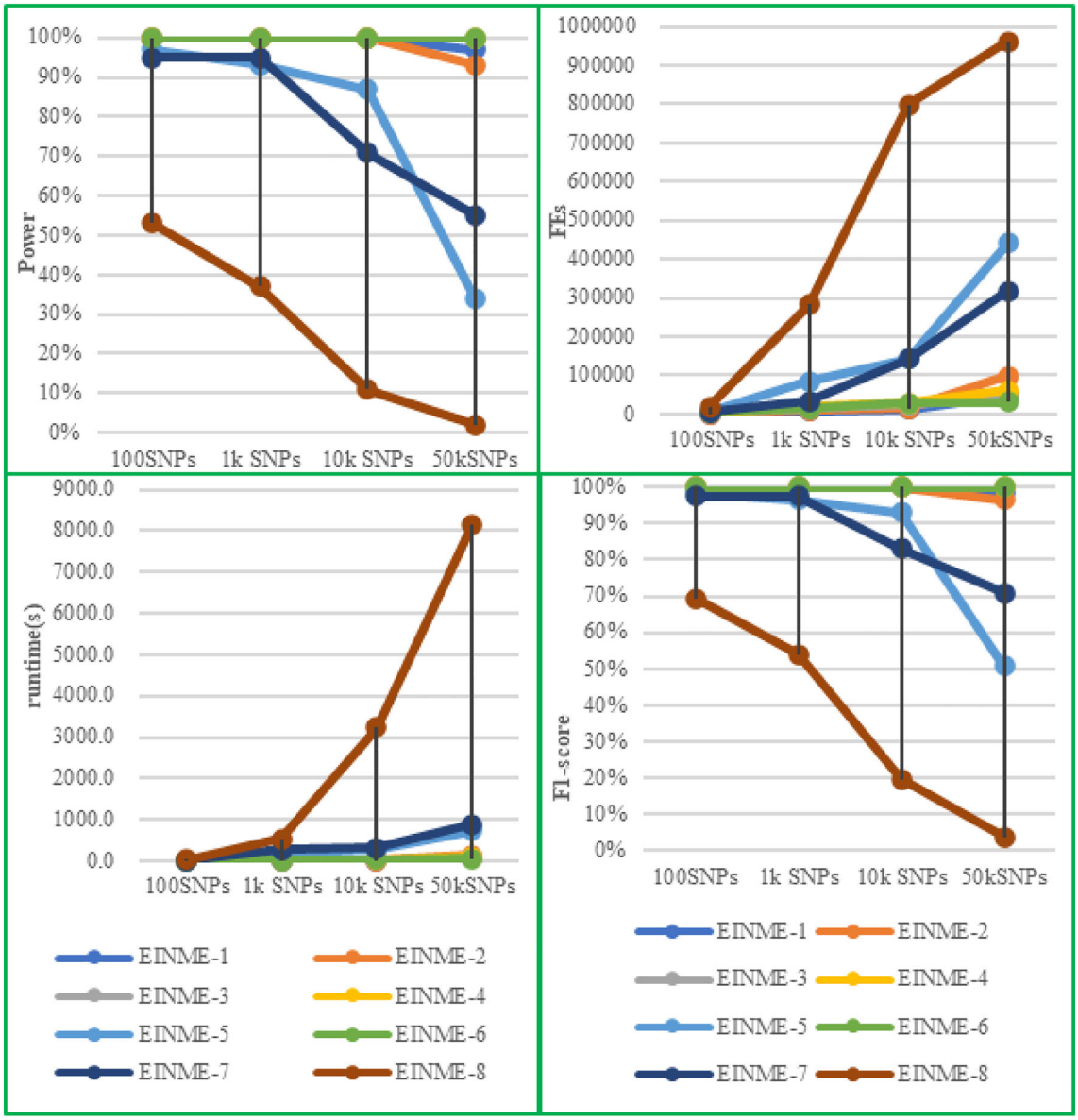

Fig. 9 Change curves of the power, FEs, runtime and F1-score of MTHSA-DHEI with an increasing number of SNPs for EINMEs

also been reported to be associated with AMD [31, 34]. In Figure S2(b), rs10272438 and rs2022251 are two important nodes, where rs2022251 (on chromosome 17, the difference p value = 0.1 between the case and control samples) has never been reported to be associated with disease.

Figure S3(a) shows the potential 4th-*order* SNP combinations with a classification accuracy greater than 85% and significance level less than 1×10^{-10}, from which it can be seen that the SNPs rs1329428, rs3922799, rs2207553, rs4585932, rs10494614, and rs967358 are central nodes, among which rs1329428 is the only SNP that has been reported. In Figure S3(b), rs10272438, rs10482918 and rs6104678 are three central nodes, where rs10482918 is in the NCAM2 gene on chromosome 21, and rs6104678, which is on chromosome 20, has been reported previously [30, 31, 68].

Figure S4(a) shows the 5th-*order* SNP interaction network, in which SNPs rs1363688, rs1329428 and rs207389 are central nodes. Figure S4(b) shows that rs10482918, rs10272438, rs1982756 and rs6104678 are central nodes. There are two 5th-order SNP combinations, namely, (rs380390, rs7322610, rs2556560, rs4689888, and rs10496217) and (rs2050733, rs207389, rs1178123, rs1329428, and rs1363688), that have a very high classification accuracy (97.9% and 95.8%, respectively) measured with MDR in the first combination, except for SNP rs380390, which shows a significant difference (p value = 6.19921E-07) between the case and control samples. The other four

Table 2 Power, FEs and runtime of the five algorithms for detecting 12 high-*order* EIMEs with 500 SNPs and 1000 SNPs. (The bold numbers indicate the optimal results)

Model	CSE			NHSA-DHSC			epiACO			MP-HS-DHSI			MTHSA-DHEI		
	Power	FEs	Run time (s)	Power	FEs	Run time (s)	Power	FEs	Run time (s)	Power	FEs	Run time (s)	Power	FEs	Run time (s)
Additive Model (500 SNPs, 5th-order)															
EIME-1	6	384,213	3106.4	50	122,863	241.6	38	33,508	87.4	99	1785	**3.9**	**100**	5029	10.5
EIME-2	45	242,509	1590.6	70	25,210	45.2	**100**	31,542	76.1	**100**	**1374**	**2.7**	**100**	4771	9.8
EIME-3	85	115,618	595.8	75	38,080	66.1	**100**	32,832	79.0	**100**	**1323**	**2.7**	**100**	5323	11.3
EIME-4	62	184,531	1362.9	74	31,259	50.3	**100**	211,316	574.4	**100**	**1371**	**2.7**	**100**	4760	9.9
Threshold Model (500 SNPs, 5th-order)															
EIME-5	80	134,314	1021.6	45	200,001	345.4	0	356,180	1015.8	89	7727	**20.0**	**100**	6263	13.2
EIME-6	66	172,331	1406.9	56	42,484	71.6	30	368,278	975.4	97	3648	**9.0**	**100**	5037	10.7
EIME-7	50	223,879	1874.9	60	40,445	71.0	47	367,892	1054.6	**100**	**1372**	**2.7**	**100**	4673	9.9
EIME-8	77	126,384	1056.1	65	37,050.5	63.3	**100**	283,818	809.4	**100**	**1368**	**2.7**	**100**	8630	19.9
Multiplicative Model (1000 SNPs, 4th-order)															
EIME-9	**100**	44,632	167.4	40	200,001	324.0	**100**	7148	33.3	93	8080	19.4	**100**	2823	**5.5**
EIME-10	70	58,060	232.9	40	200,001	309.6	2	177,132	848.6	98	3046	6.6	**100**	2694	**5.3**
EIME-11	35	73,662	328.9	85	12,727	18.2	79	8912	46.1	99	2600	5.4	**100**	2690	**5.3**
EIME-12	**100**	44,663	181.3	**100**	4492	**6.2**	**100**	5608	29.5	**100**	3068	6.4	**100**	2559	**4.8**

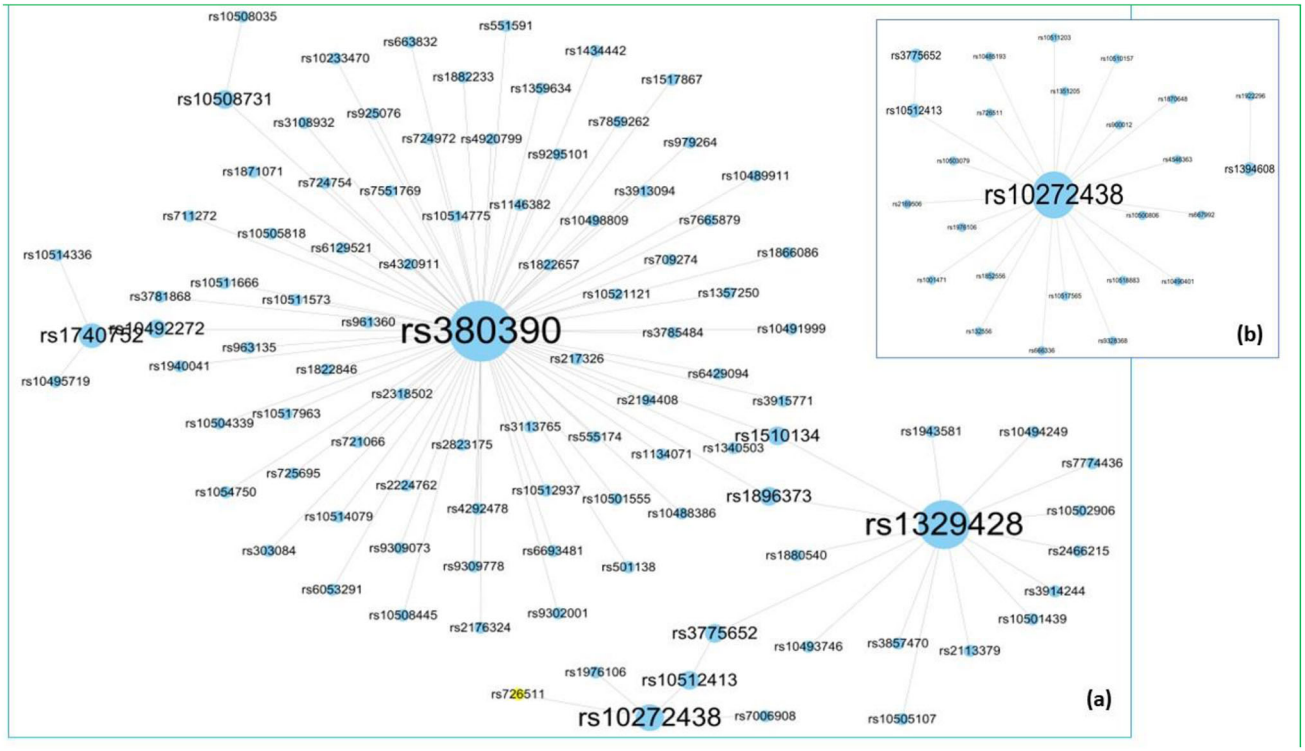

Fig. 10 A 2nd-*order* SNP interaction network. A node denotes a SNP locus, and an edge indicates that the 2nd-*order* SNP combination connected by that edge is strongly associated with AMD (the significance level for the G-test is $< 1 \times 10^{-8}$, and the classification accuracy of the 2nd-*order* SNP combination is > 75%). The larger the node is, the greater the number of nodes connected to it

SNPs (rs7322610, rs2556560, rs4689888, and rs10496217) have very small effect sizes, and their p values are equal to 0.282, 0.283, 0.864 and 0.220, respectively. For the 2nd SNP combination, the p values of the five SNPs were equal to 0.324, 0.088, 0.011, $5.99E-06$ and $3.84E-05$.

To complete detection with 5×10^7 FEs, the proposed MTHSA-DHEI method took no more than 20 h to simultaneously detect 2nd-*order*, 3rd-*order*, 4th-*order* and 5th-*order* SNP epistatic interactions from the AMD dataset. The time required for MP-HS-DHSI to perform the detection of 2nd-*order*, …, 5th-*order* SNP epistatic interactions one by one was more than 48 h; NHSA-DHSC, CSE and epiACO required more than one days. According to the AMD detection results, MTHSA-DHEI found almost all the SNPs that have been reported to be associated with AMD, such as rs380390, rs1329428, rs1363688, rs10272438, and rs555174, and found some SNPs that have been reported to be associated with other complex diseases. Some previously unreported SNPs were also found by MTHSA-DHEI and are worthy of further study by biologists.

Conclusion

According to the experimental results, the proposed MTHSA-DHEI method is significantly superior to the other four algorithms in terms of power, FEs and runtime for the EINMEs. The EIMEs also outperform others with respect to power and FEs, but they take more time than MP-HS-DHEI for most EIMEs. In the AMD experiment, MTHSA-DHEI also showed a powerful ability to detect high-*order* epistatic interactions from hundreds of thousands of data points, and it found almost all the SNPs that have been reported to be associated with AMD. Although the results of simulation experiments indicate that our method outperforms the four compared SISAs and shows very effective performance for detecting high-order SNP epistatic models, such as EINME-1, EINME-4, and EINME-6, additive models and threshold models, it still cannot ensure the identification of casual epistatic interactions from a dataset of over 10,000 SNPs in a limited amount of time (30 min), and detection power starts to degrade rapidly, such as for EINME-8. For the four multiplicative models, the heritability and population prevalence values have a very important influence on the detection power of the algorithm. The larger the heritability and population prevalence values of the disease model are, the higher the

detection power of the algorithm. However, SNP loci typically have a small effect and only modest heritability [2, 69]. To enhance the detection power of the proposed MTHSA-DHEI method on datasets with more than 10,000 SNPs, we can set a large size for HM and EHS and set a large value for MaxFEs to make the algorithm run for a longer time. Large sizes of harmony memory and large values of MaxFEs for our algorithm can improve detection power, but the computational burden will also increase rapidly.

Discussion

Traditional methods for detecting high-*order* SNP epistatic interactions can perform only a single task and ignore the sharing of information between tasks, which makes the computational burden of detecting SNP epistatic interactions from unknown disease data very high. To address this problem, this study aimed to improve detection power, reduce the computational burden and enhance the ability to discriminate high-*order* SNP epistatic interactions from a significant number of high-*order* SNP combinations. We proposed a novel multitasking HS algorithm for detecting high-*order* SNP epistatic interactions, where multitasking is applied to accelerate detection using concurrent collaborative computation, transfer learning is adopted to enhance the information exchange between tasks, and four complementary evaluation functions are employed to promote the ability to identify various disease models and overcome the preference of a single evaluation function for a specific disease model. In addition, for the epistatic interaction model with no marginal effects, it is very difficult to uncover clues that can guide the search algorithm to locate the functional SNP loci. The proposed MTHSA-DHEI algorithm integrates ND-JE into the evaluation functions to seek clues of the functional SNP locus that has no or a very weak marginal effect on disease status.

MTHSA-DHEI is a metaheuristic search algorithm, and its time complexity is determined by four objective functions (K2-score, ME score, LR score and ND-JE-score) of the 1st stage and the MaxFEs (maximum number of evaluations of associations between SNP combinations and disease status). In the 2nd stage and 3rd stage, only the G-test and MDR are employed to test and verify the number of candidate solutions that were found in the 1st stage, and the associated time complexity is negligible. The time complexity of evaluating objective functions is $O(k \times S)$ (where k is the order of SNP combinations and S is the number of samples). Therefore, the time complexity of MTHSA-DHEI is roughly equivalent to $O(k \times S \times \text{MaxFEs})$, which is much less than the time complexity $O(k \times S \times N^k)$ of the exhaustive method, where N is the number of SNPs in the dataset. Since N and MaxFEs are much larger than k and S, the time complexity of the traditional exhaustive method is $O(N^k)$, which is much higher than the time complexity O(k × S × MaxFEs) of MTHSA-DHEI.

To the best of our knowledge, this study is the first to detect high-*order* SNP epistatic interactions by using a multitasking search algorithm. There is still much room for improving the performance of this type of algorithm. In the future, we should try to develop an explicit-encoding-based multitasking search algorithm to improve the search speed and design more effective evaluation functions for identifying various disease models.

In addition, the fuzzy set-based optimization algorithm [70] has received much attention in recent years and has been successfully applied to solid assignment [71, 72] and transportation [73] problems, and it can also be considered a focal method for future studies on high-order SNP epistatic interaction detection. In addition, it is also very important to develop an effective scoring function for seeking clues to guide the SISA to locate the positions of potential SNP interactions. The proposed MTHSA-DHEI method can only be applied to detect associations between common SNPs (MAF > 0.05 && MAF < 0.5) and disease status. This needs to be further studied for its application to rare variants.

Availability and implementation

The supplementary files, MATLAB codes and Python code are available at https://github.com/shouhengtuo/MTHSADHEI.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40747-022-00813-7.

Acknowledgements The author would like to thank all the editors, reviewers and referees for their constructive comments.

Funding This research was supported by Natural Science Foundation of China (Grant 62002289).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Guo X (2015) Searching genome-wide disease association through SNP Data. Dissertation, Georgia State University. https://scholarworks.gsu.edu/cs_diss/101.
2. Manolio TA et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–753
3. Easton DF et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087–1093
4. Fellay J et al (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317:944–947
5. Wang MH, Cordell HJ, Van Steen K (2019) Statistical methods for genome-wide association studies. Semin Cancer Biol 55:53–60
6. Visscher PM, Wray NR, Zhang Q et al (2017) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101:5–22
7. Upton A, Trelles O, Cornejo-Garcia JA, Perkins JR (2016) Review: high-performance computing to detect epistasis in genome scale datasets. Brief Bioinform 17(3):368–379. https://doi.org/10.1093/bib/bbv058
8. Loucoubar C, Grant AV, Bureau J-F et al (2017) Detecting multiway epistasis in family-based association studies. Brief Bioinform 18(3):394–402. https://doi.org/10.1093/bib/bbw039
9. Li P, Guo MZ, Wang CY et al (2015) An overview of SNP interactions in genome-wide association studies. Brief Funct Genomics 14:143–155
10. Banerjee S, Zeng LY, Schunkert H et al (2018) Bayesian multiple logistic regression for case–control GWAS. PLoS Genet 14:27
11. Sun S, Dong B, Zou Q (2021) Revisiting genome-wide association studies from statistical modelling to machine learning. Brief Bioinform 22(4):263. https://doi.org/10.1093/bib/bbaa263
12. Gros PA, Le Nagard H, Tenaillon O (2009) The evolution of epistasis and its links with genetic robustness, complexity and drift in a phenotypic model of adaptation. Genetics 182(1):277–293. https://doi.org/10.1534/genetics.108.099127
13. Zhang Y, Liu J (2007) Bayesian inference of epistatic interactions in case–control studies. Nat Genet 39:1167–1173. https://doi.org/10.1038/ng2110
14. Guo X, Meng Y, Yu N, Pan Y (2014) Cloud computing for detecting high *order* genome-wide epistatic interaction via dynamic clustering. BMC Bioinformatic 5(1):102
15. Yang GYJW, Yang Q et al (2014) PBOOST: a GPU-based tool for parallel permutation tests in genome-wide association studies. Bioinformatics 2014(9):1460–1462
16. Cecilia JM, Ponte-Fernández C, González-Domínguez J, Martín MJ (2020) Fast search of third-*order* epistatic interactions on CPU and GPU clusters. Int J High Perform Comput Appl 34(1):20–29. https://doi.org/10.1177/1094342019852128
17. Wang J, Joshi T, Valliyodan B, Shi H, Liang Y et al (2015) A Bayesian model for detection of high-*order* interactions among genetic variants in genome-wide association studies. BMC Genomics 16:1011. https://doi.org/10.1186/s12864-015-2217-6
18. Han B, Chen XW, Talebizadeh Z, Xu H (2012) Genetic studies of complex human diseases: characterizing SNP-disease associations using Bayesian networks. BMC Syst Biol 6(Suppl 3):S14. https://doi.org/10.1186/1752-0509-6-S3-S14
19. Wang W (2010) TEAM: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics 26(12):i217
20. Moore JH, Hahn LW, Ritchie MD, Thornton TA, White BC (2002) Application of genetic algorithms to the discovery of complex genetic models for simulation studies in human genetics. In: Langdon WB, et al., editors. Proceedings of the Genetic and Evolutionary Computation Conference. Morgan Kaufmann Publishers; San Francisco
21. Moore JH, Hahn LW, Ritchie MD et al (2004) Routine discovery of complex genetic models using genetic algorithms. Appl Soft Comput 4(1):79–86
22. Moore JH, Andrews PC, Olson RS, Carlson SE, Larock CR, Bulhoes MJ, Armentrout SL (2017) Grid-based stochastic search for hierarchical gene–gene interactions in population-based genetic studies of common human diseases. BioData Mining 10:19. https://doi.org/10.1186/s13040-017-0139-3
23. Wang Y, Liu X, Robbins K et al (2010) AntEpiSeeker: detecting epistatic interactions for case–control studies using a two-stage ant colony optimization algorithm. BMC Res Notes 3(1):117
24. Shang J, Zhang J, Lei X, Zhang Y, Chen B (2012) Incorporating heuristic information into ant colony optimization for epistasis detection. Genes Genom 34(3):321–327
25. Sun Y, Shang J, Liu JX, Li S, Zheng CH (2017) epiACO—a method for identifying epistasis based on ant Colony optimization algorithm. BioData Mining 10:23. https://doi.org/10.1186/s13040-017-0143-7
26. Sun Y, Wang X, Shang J, Liu J, Zheng C, Lei X (2019) Introducing heuristic information into ant colony optimization algorithm for identifying epistasis. IEEE/ACM Trans Comput Biol Bioinform. https://doi.org/10.1109/TCBB.2018.2879673
27. Yang CH, Chuang LY, Lin YD (2017) Multi-objective differential evolution-based multifactor dimensionality reduction for detecting gene–gene interactions. Sci Rep 7(1):12869. https://doi.org/10.1038/s41598-017-12773-x
28. Yang CH, Kao YK, Chuang LY, Lin YD (2018) Catfish taguchi-based binary differential evolution algorithm for analysing single nucleotide polymorphism interactions in chronic dialysis. IEEE Trans Nanobiosci 17(3):291–299
29. Aflakparast M et al (2014) Cuckoo search epitasis: a new method for exploring significant genetic interactions. Heredity 112:666–674
30. Tuo S, Zhang J, Yuan X et al (2016) FHSA-SED: two-locus model detection for genome-wide association study with harmony search algorithm. PLoS One 11(3):e0150669
31. Tuo S, Zhang J, Yuan X, He Z, Liu Y, Liu Z (2017) Niche harmony search algorithm for detecting complex disease associated high-*order* SNP combinations. Sci Rep 7:11529
32. Shouheng T, Haiyan L, Hao C (2020) Multipopulation harmony search algorithm for the detection of high-*order* SNP interactions. Bioinformatics 36:4389–4398. https://doi.org/10.1093/bioinformatics/btaa215
33. Wang J, Joshi T, Valliyodan B, Shi H, Liang Y, Nguyen HT et al (2015) A Bayesian model for detection of high-*order* interactions among genetic variants in genome-wide association studies. BMC Genomics 16:1011. https://doi.org/10.1186/s12864-015-2217-6
34. Guo Y, Zhong Z, Yang C, Hu J, Jiang Y, Liang Z et al (2019) Epi-GTBN: an approach of epistasis mining based on genetic Tabu algorithm and Bayesian network. BMC Bioinform 20(1):444. https://doi.org/10.1186/s12859-019-3022-z
35. Visweswaran S, Wong AKI, Barmada MM (2009) A Bayesian method for identifying genetic interactions[C]. AMIA Ann Sympos Proc Am Med Inform Assoc: 673
36. Cao X, Yu G, Liu J, Jia L, Wang J (2018) ClusterMI: detecting high-*Order* SNP interactions based on clustering and mutual information. Int J Mol Sci 19(8):2267
37. Jing PJ, Shen HB (2015) MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide

association studies. Bioinformatics 31:634–641. https://doi.org/10.1093/bioinformatics/btu702
38. Crawford L, Zeng P, Mukherjee S, Zhou X (2017) Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet 13(7):e1006869. https://doi.org/10.1371/journal.pgen.1006869
39. Gola D, Mahachie John JM, van Steen K, König IR (2016) A roadmap to multifactor dimensionality reduction methods. Brief Bioinform 17(2):293–308. https://doi.org/10.1093/bib/bbv038
40. Kim H, Jeong HB, Jung HY, Park T, Park M (2019) Multivariate cluster-based multifactor dimensionality reduction to identify genetic interactions for multiple quantitative phenotypes. Biomed Res Int 2019:4578983. https://doi.org/10.1155/2019/4578983
41. Gupta A, Ong YS, Feng L (2016) Multifactorial evolution: toward-stoward evolutionary multitasking. IEEE Trans Evol Comput 20(3):343–357
42. Tang ZD, Gong MG et al (2021) A multifactorial optimization framework based on adaptive intertask coordinate system. IEEE Trans Cybernet. https://doi.org/10.1109/TCYB.2020.3043509
43. Li JZ, Li H et al (2021) Multi-fidelity evolutionary multitasking optimization for hyperspectral endmember extraction. Appl Soft Comput 111:107713
44. Feng L et al (2019) Explicit evolutionary multitasking for combinatorial optimization: a case study on capacitated vehicle routing problem. IEEE Trans Cybernet 51(6):3143–3156. https://doi.org/10.1109/TCYB.2019.2962865
45. Osaba E, Del Ser J, Martinez AD, Lobo JL, Herrera F (2021) AT-MFCGA: an adaptive transfer-guided multifactorial cellular genetic algorithm for evolutionary multitasking. Inf Sci 570:577–598
46. Tam NT, Dat VT, Lan PN, Binh HTT, Vinh LT, Swami A (2021) Multifactorial evolutionary optimization to maximize lifetime of wireless sensor network. Inf Sci 576:355–373
47. Xu X, Yin G, Wang C (2021) Multitasking scheduling with batch distribution and due date assignment. Complex Intell Syst 7:191–202. https://doi.org/10.1007/s40747-020-00184-x
48. Dang Q, Gao W, Gong M (2022) Multi-objective multitasking optimization assisted by multidirectional prediction method. Complex Intell Syst. https://doi.org/10.1007/s40747-021-00624-2
49. Zhao Y, Ye S, Chen X et al (2021) Polynomial Response Surface based on basis function selection by multitask optimization and ensemble modeling. Complex Intell Syst. https://doi.org/10.1007/s40747-021-00568-7
50. Neapolitan RE (2004) Learning bayesian networks. Prentice Hall, Upper Saddle River
51. Li X (2017) A fast and exhaustive method for heterogeneity and epistasis analysis based on multi-objective optimization. Bioinformatics 18:2829–2836. https://doi.org/10.1093/bioinformatics/btx339
52. Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD (2008) Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. BMC Bioinform 9:238. https://doi.org/10.1186/1471-2105-9-238
53. Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: part 1. Biometrika 20A:175–240
54. Geem ZW, Kim JH, Loganathan GV (2001) A new heuristic optimization algorithm: harmony search. SIMULATION 76(2):60–68
55. Das S, Mukhopadhyay A, Roy A, Abraham A, Panigrahi BK (2011) Exploratory power of the harmony search algorithm: analysis and improvements for global numerical optimization. Syst Man Cybernet Part B 41(1):89–106
56. Tuo S, Geem ZW, Yoon JH (2020) A new method for analyzing the performance of the harmony search algorithm. Mathematics 8(9):1421. https://doi.org/10.3390/math8091421
57. Zhang TH, Geem ZW (2019) Review of harmony search with respect to algorithm structure. Swarm Evol Comput 48:31–43
58. Crow Jf (1999) Hardy. Weinberg and language impediments. Genetics 152:821–825
59. Hoey J (2012) The two-way likelihood ratio (G) test and comparison to two-way chi squared test. arXiv preprint arXiv:1206.4881
60. Himmelstein et al (2011) Evolving hard problems: generating human genetics datasets with a complex etiology. BioData Min. https://doi.org/10.1186/1756-0381-4-21
61. Ponte-Fernández C, González-Domínguez J, Carvajal-Rodríguez A et al (2020) Toxo: a library for calculating penetrance tables of high-*order* epistasis models. BMC Bioinform. https://doi.org/10.1186/s12859-020-3456-3
62. Urbanowicz RJ, Kiralis J, Sinnott-Armstrong NA, Heberling T, Fisher JM, Moore JH (2012) GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData mining 5:1–14
63. Klein RJ et al (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389
64. Xie M, Li J, Jiang T (2012) Detecting genome-wide epistasis based on the clustering of relatively frequent items. Bioinformatics 28(1):5–12. https://doi.org/10.1093/bioinformatics/btr603
65. Barba M, Pietro LD, Massimi L et al (2018) BBS9 gene in nonsyndromic craniosynostosis: Role of the primary cilium in the aberrant ossification of the suture osteogenic niche. Bone 112:58–70
66. Mirabello L, Richards EG, Duong LM et al (2011) Telomere length and variation in telomere biology genes in individuals with osteosarcoma. Int J Mol Epidemiol Genet 2(1):19–29
67. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–504. https://cytoscape.org/
68. Jiang R, Tang W, Wu X, Fu W (2009) A random forest approach to the detection of epistatic interactions in case–control studies. BMC Bioinform 10(Suppl 1):S65. https://doi.org/10.1186/1471-2105-10-S1-S65
69. Tam V, Patel N, Turcotte M et al (2019) Benefits and limitations of genome-wide association studies. Nat Rev Genet 20:467–484. https://doi.org/10.1038/s41576-019-0127-1
70. Kumar PS (2020) Algorithms for solving the optimization problems using fuzzy and intuitionistic fuzzy set. Int J Syst Assur Eng Manag 11(1):189–222. https://doi.org/10.1007/s13198-019-00941-3
71. Kumar PS (2019) Intuitionistic fuzzy solid assignment problems: a software-based approach. Int J Syst Assur Eng Manag 10(4):661–675. https://doi.org/10.1007/s13198-019-00794-w
72. Kumar PS (2020) The PSK method for solving fully intuitionistic fuzzy assignment problems with some software tools. Adv Bus Strategy Compet Adv. https://doi.org/10.4018/978-1-5225-8458-2.ch009
73. Kumar PS (2021) Finding the solution of balanced and unbalanced intuitionistic fuzzy transportation problems by using different methods with some software packages. Handbook Res Appl AI Int Bus Market Appl. https://doi.org/10.4018/978-1-7998-5077-9.ch015