Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(6): 7711–7726.
DOI: 10.3934/mbe.2021382
Received: 25 May 2021
Accepted: 01 September 2021
Published: 07 September 2021

Research article

An epistasis and heterogeneity analysis method based on maximum correlation and maximum consistence criteria

Xia Chen[1,2], Yexiong Lin[2], Qiang Qu[2], Bin Ning[2], Haowen Chen[2,]* and Xiong Li[3]

[1] School of Basic Education, Changsha Aeronautical Vocational and Technical College, Changsha, Hunan 410124, China
[2] College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan 410082, China
[3] School of Software, East China Jiaotong University, Nanchang 330013, China

* **Correspondence:** Email: hwchen@hnu.edu.cn; Tel: +86-073188821060; Fax: +86073188821060.

Abstract: Tumor heterogeneity significantly increases the difficulty of tumor treatment. The same drugs and treatment methods have different effects on different tumor subtypes. Therefore, tumor heterogeneity is one of the main sources of poor prognosis, recurrence and metastasis. At present, there have been some computational methods to study tumor heterogeneity from the level of genome, transcriptome, and histology, but these methods still have certain limitations. In this study, we proposed an epistasis and heterogeneity analysis method based on genomic single nucleotide polymorphism (SNP) data. First of all, a maximum correlation and maximum consistence criteria was designed based on Bayesian network score *K2* and information entropy for evaluating genomic epistasis. As the number of SNPs increases, the epistasis combination space increases sharply, resulting in a combination explosion phenomenon. Therefore, we next use an improved genetic algorithm to search the SNP epistatic combination space for identifying potential feasible epistasis solutions. Multiple epistasis solutions represent different pathogenic gene combinations, which may lead to different tumor subtypes, that is, heterogeneity. Finally, the XGBoost classifier is trained with feature SNPs selected that constitute multiple sets of epistatic solutions to verify that considering tumor heterogeneity is beneficial to improve the accuracy of tumor subtype prediction. In order to demonstrate the effectiveness of our method, the power of multiple epistatic recognition and the accuracy of tumor subtype classification measures are evaluated. Extensive simulation results show that our method has better power and prediction accuracy than previous methods.

Keywords: genetic algorithm; Bayesian network; information entropy; tumor subtype classification;

genome variation

1. Introduction

Precision medicine is the most potential strategy to improve the clinical treatment of cancer. In-depth study of tumor heterogeneity is an important prerequisite for the implementation of precision medicine [1–3]. Tumor heterogeneity can be divided into two categories: inter-tumor heterogeneity and intra-tumor heterogeneity. The former refers to the molecular and cellular differences in the same organ tumors of different tumor patients, and the latter refers to the molecular and cellular differences in different tumor formation sites of the tumor tissue of the same patient [4–7]. Due to tumor heterogeneity, the same drugs and treatment measures have obvious clinical efficacy differences for different tumor subtypes [8–10]. For intra-tumor heterogeneity, there may be different tumor cell subclones in the same tumor tissue. When the drug only has a therapeutic effect on the primary clonal tumor cell, the remaining secondary subclonal cells will further evolve, leading to tumor recurrence and metastasis. More importantly, different tumor subtypes or different subclonal tumor cells may adopt different immune escape mechanisms, so that identifying their immune escape pathways in a targeted manner can implement immunotherapy more effectively [11–14]. In short, it is necessary to in-depth study tumor heterogeneity.

Fortunately, with the help of next-generation high-throughput sequencing technology, researchers can use different levels of omics data (e.g., genome [15–18], transcriptome [19–22], proteome [23,24], metabolome and epitome [25–27], etc.) to stratify tumor cohorts accurately. On the basis of accurately identifying the patient's tumor subtype, precision medicine can be implemented for the patient. Turashvili et al. summarized a comprehensive survey on inter-tumor heterogeneity and intra-tumor heterogeneity aspects for breast cancer [28]. For example, Hofree et al. proposed a network-based hierarchical method NBS which first converted somatic mutation data at the genome level into continuous analog signals through a network smoothing method, and then used a non-negative matrix factorization method to identify patient samples. In order to improve the robustness of the clusters, the consensus clustering algorithm was used to determine the number of clusters and the cluster relationship of samples [29]. For genome-level single nucleotide polymorphism (SNP) data, Li et al. proposed a three-stage processing framework based on techniques such as multi-objective optimization algorithm [30–33], clustering algorithm [34] and deep learning [35–38], which dealt with the problems of epistasis, heterogeneity [38–40] and tumor subtype prediction [38–43] respectively, which improved the efficiency of epistasis analysis and the accuracy of tumor heterogeneity recognition to a certain extent. Based on transcriptome data, Jiang et al. divided triple negative breast cancer (TNBC) into four subtypes and putative therapeutic targets or biomarkers were identified for each subtype [44]. To recognize dominate evasion pathway of different subtypes of breast cancer, Bou-Dargham et al. collected 1356 immune-related genes as clustering features based on cohorts' transcriptome data [45]. However, the occurrence and development of tumors usually involve a variety of biomolecules, and only a certain level of omics data is difficult to fully characterize its internal characteristics, resulting in difficult reproducibility, lack of interpretability, and so on. Therefore, with integrating multiple omics data such as genome, epigenome and transcriptome, Robertson et al. typed the 80 cases of uveal melanoma samples, and four subtypes with significant molecular differences and clinically relevant were identified [46]. More importantly, the role of single-cell sequencing data in tumor heterogeneity research has received widespread attention. For instance, Xiong et al. compared the heterogeneity among patients and subtypes based on CNV and revealed that there are four subtypes in Glioblastoma cells based on single-cell analysis [47]. Lawson et al. summarized cellular differentiation, diagnostics and therapy response, metastasis and heterogeneity in the microenvironment at single-cell resolution [48].

With the accumulation of genomic variation, different subclones are differentiated in tumor cells, each of which may be caused by different epistasis of pathogenic genes. Our research designs a new risk epistatic combination evaluation method, and then improves the heuristic genetic algorithm to search the epistasis combination space to overcome the combination explosion challenge. The multiple risk epistasis combinations identified by our method represent tumor heterogeneity. Using feature SNPs to train the XGBoostclassifier, the results of subtype classification show that accurately identifying tumor heterogeneity is beneficial to improve the performance of tumor subtype classification.

2. Materials and method

This study mainly faces two challenges. One is that with the increase in the number of SNPs, the combination space of high-order epistasis increases sharply, which leads to a combination explosion. The other is that the epistasis between susceptible genes and the heterogeneity among tumor samples need to be considered. In order to solve these two problems, we first design multiple objectives evaluation criteria to evaluate the epistasis combination of susceptible genes from different angles and improves the heuristic genetic algorithm to efficiently search the combination space to screen multiple candidate epistasis combinations.

2.1. Maximum correlation and maximum consistence criteria

A single optimization criterion can usually only find potential risk epistatic combinations from one angle, which may miss other feasible solutions, especially when there is heterogeneity in the sample [49]. In this study, two evaluation criteria are designed to evaluate candidate epistasis combinations. One is to examine the correlation between the epistasis combination and the target phenotype, and the other is to examine the rationality of genotypes distribution of the epistasis combination among the samples. These two goals have evaluated the superior combination from different angles and have a certain degree of complementarity.

Maximum correlation: It can be assumed that if a combination of loci is pathogenic, then the genotypes at these loci should have a strong correlation with the case/control states of the samples. Therefore, we can first use the maximum correlation to quantify and rank each candidate epistasis combination. Bayesian network consisted of nodes and directed edges can be used to describe the relationship between epistatic SNPs and disease. Given the Markov condition, the relationship between k SNPs and sample disease state can be simplified as the calculation of Eq (1), namely the joint probability distribution of epistatic combination in training samples.

$$p(X_1, X_2, ..., X_{k+1}) = \prod_{i=1}^{k+1} P(X_i \mid \pi(X_i)) \tag{1}$$

where $\pi(X_i)$ is the parents nodes of the node X$_i$ and if $\pi(X_i) = \varnothing$, $P(X_i \mid \pi(X_i))$ equals to a marginal distributions. *K2* is a measure score for evaluating the structure of Bayesian network and it has been widely used in previous studies [30,50]. Since there is no prior knowledge about the known pathegenic SNPs in this study, the Dirichlet distribution $D[\alpha_{11}...\alpha_{ij}]$ are set to be 1 in Eq (2).

$$K2 = \sum_{i=1}^{I}\left(\sum_{b=1}^{r_i+1}\log(b) - \sum_{j=1}^{2}\sum_{d=1}^{r_{ij}}\log(d)\right) \tag{2}$$

where *I* is the combination space of *k* SNPs and each SNP has 3 states, namely heterozygous, wild homozygous and mutant homozygous, so that $I = 3^k$. r_i denotes the frequency of *i*-th epistatic combination in training samples and r_{ij} represents the frequency of *i*-th epistatic combination in samples with *j*-th state and j ∈(case, control).

Maximum consistence: Due to the heterogeneity in the samples, there may not be a strong correlation between some epistatic combinations and sample labels. Therefore, it is necessary to find these potentially susceptible genotypes from another perspective. It can be assumed that there is another susceptible genotype in the sample due to heterogeneity, and this genotype has a weaker association due to its small sample size. Therefore, this study believes that it can be investigated from the perspective of genotype consistency. If the genotype at the loci is more stable in the sample, then the genotype may be one of the main sources of case samples. Here, we applied Shannon entropy to measuring the uncertainty of genotype distribution. Of note, the smaller the information entropy value, the more stable the genotype at the loci is in the case samples, and the greater the possibility that the genotype is risky. The uncertainty of genotype distribution is defined as Eq (3).

$$HE = -\sum_{i=1}^{h} p_i \log p_i \tag{3}$$

where *h* refers to the number of genotypes actually present in the case samples and p_i is the possibility of the *i*-th genotype in the case samples.

2.2. Adaptive genetic algorithm

The genetic algorithm simulates the phenomena of replication, crossover, and mutation that occur in natural evolution. Starting from the initial population, through random selection, crossover and mutation operations, a group of individuals (solution) more adapted to the environment (problem) are generated and finally converges to a group of individuals who are most adaptable to the environment. The genetic algorithm includes several main parts: coding, fitness value evaluation, selection, crossover, and mutation, which are briefly introduced below.

Individual coding: Individual coding is the first step in using genetic algorithms to solve specific problems. We encode each individual as a string of ‘0’, ‘1’, and the length of the string is equal to the number of SNPs, ‘0’ means that the SNP is not selected, and ‘1’ means that the SNP is selected to construct an epistatic combination. The goal of our study is to identify epistasis and heterogeneity. Therefore, multiple ‘1’s on an individual solution indicate epistasis, and multiple near-optimal solutions correspond to heterogeneity.

Fitness evaluation: The evaluation of the solution not only affects the convergence speed of the population, but also determines whether to break out of the local optimum. Here, we combined the maximum correlation and maximum consistence criteria to evaluate each solution.

Selection operator: The selection operation simulates the rule of survival of the fittest in the natural evolution process. This operator keeps the more adaptable individuals in the population and recombines them to produce better next generations. In this study, each candidate solution is investigated with two criteria, and the sufficiently good solutions on any one goal are retained for the next generation.

Crossover operator: The size of the crossover probability P_c determines the abundance of the population. The larger the Pc, the higher the abundance of the population, but the higher the probability that good individuals will be destroyed [51]. In this study, we design an adaptive crossover probability adjustment strategy defined as Eq (4) to dynamically adjust based on the optimal fitness value and average fitness value of the population.

$$P_c = \begin{cases} k_1 (f_{\max} - \dfrac{f_1 + f_2}{2}) / (f_{\max} - f_{ave}) & \dfrac{f_1 + f_2}{2} \geq f_{ave} \\ k_2 & else \end{cases} \tag{4}$$

where $f_{\max}$ is the maximum fitness of the population and f_{ave} is the average fitness of the population. The f_1 and f_2 are the fitness of two individuals to be recombined. The k_1 and k_2 are constant values between 0 and 1.

Mutate operator: The size of the mutation probability P_m defined as Eq (5) determines whether it can jump out of the local optimum value to find the global optimal solution. The larger the P_m, the easier it is to jump out of the local optimum value to find the global optimal solution. However, a too large P_m value will degenerate genetic algorithm to Random search.

$$P_m = \begin{cases} k_3 (f_{\max} - f) / (f_{\max} - f_{ave}) & f \geq f_{ave} \\ k_4 & else \end{cases} \tag{5}$$

where f is the fitness of the individual to be mutated and k_3 and k_4 are constant values between 0 and 1.

2.3. Tumor subtype recognition and classification

In this section, K-means clustering algorithm was introduced to recognize different tumor subtypes hidden in cases. For a given sample set, K-means divides the sample set into *K* clusters according to the distance between the samples. Make the points in the clusters as close together as possible, and make the distance between the clusters as large as possible. Assuming that the cases are divided into *K* clusters, namely ($C1$, $C2$, ... Ck), then our goal is to minimize the squared error *E* defined in Eq (6). Of note, the *K* is set to be number of risky epistatic combinations, which means that each risk epistasis may lead to a subtype.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|_2^2 \tag{6}$$

where μ_i is the centroid of i-th cluster, and is defined as Eq (7). The optimization objective can be solved by iterative algorithms.

$$\mu_i = \frac{1}{|C_i|}\sum_{x\in C_i} x \tag{7}$$

After using the K-means clustering algorithm to subdivide all case samples into different subtypes, our method proposes to use the XGBoost classifier to predict the subtype of a new sample. The XGBoost implements parallel tree boosting technique in a portable and efficient way [52]. The XGBoost prevents over-fitting through regularization items, and it can be optimized in parallel to improve training efficiency. To help researchers to solve classification problem in an easy-use, friendly way, R, Python, Ruby are provided. In order to more intuitively describe the computing framework proposed in this paper, we draw a flow chart in Figure 1 to describe the method in this paper.

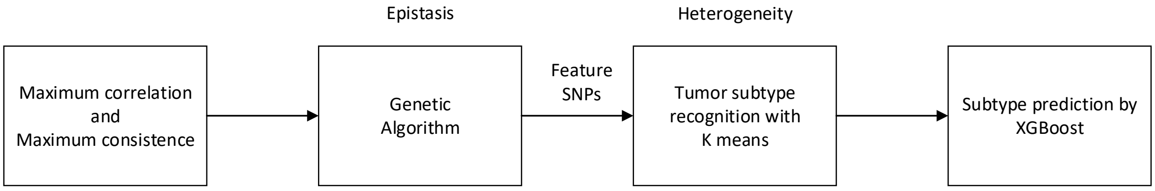
Epistasis
Heterogeneity
Maximum correlation and Maximum consistence
Genetic Algorithm
Feature SNPs
Tumor subtype recognition with K means
Subtype prediction by XGBoost

Figure 1. The flowchart of our method for epistasis and heterogeneity analysis.

3. Datasets and evaluation measures

3.1. Simulation datasets

In this study, to demonstrate the performance of the proposed method, extensive simulated datasets listed in Table 1 were generated by GAMETES_2.1 [53] which has been wildly used in previous studies [34,35]. With using GAMETES_2.1, researchers can customize epistatic datasets by setting parameters, such as minor allele frequency (MAF), epistatic order (k), heterogeneity proportion (HP), sample size, SNP size and so on.

In this study, we also simulated pure and heterogeneous dataset with using the same parameters as previous study [35]. The pure datasets were labelled with ‘Pure’ prefix, while heterogeneous datasets were labelled with ‘Hete’ prefix. The sample sizes of the data sets have five scales, namely 1000, 2000, 3000, 4000 and 8000, but the number of SNPs in each data set is fixed at 100. For pure datasets (HP = 1.0), 2 loci and 3 loci disease models were simulated and their MAFs are (0.2, 0.2) and (0.2, 0.2, 0.2) and their epistasis order are 2 and 3, respectively. For heterogeneous datasets, all these datasets are composed with two balanced disease models, so that the disease model H1 holds 50% and the H2 is also 50%. For instance, the HPs of ‘Hete10’ dataset are H1 = 50% and H2 = 50%, so that the 8000 samples are composed with 4000 samples caused by 3 loci disease model H1(0.2, 0.2, 0.2) and the rest 4000 samples caused by 3 loci disease model H2 (0.3,0.3,0.3).

3.2. Performance evaluation measures

This study has dealt with the epistasis and heterogeneity of complex diseases at the same time, so different measures are used to evaluate the performance of our method. The Power is defined as Eq (8).

$$Power = \frac{n}{N} \tag{8}$$

where n refers the frequency of correctly identifying the real pathogenic loci and N is the number of tests.

Accurately identifying heterogeneity in cases is a prerequisite for improving the accuracy of tumor subtype classification. In this study, the accuracy defined as Eq (9) is also applied to evaluating the performance of tumor subtype classification. In order to evaluate the accuracy of tumor subtype classification more objectively, this study adopted a 10-fold cross-validation strategy which divides the dataset into 10 equal parts, and then uses each part as the test set in turn, and the remaining parts as the training set and finally the results of the 10 tests are averaged [25,54–56].

$$Acc = \frac{n}{N} \tag{9}$$

where N is the total number of cases and n represents the number of times the sample was correctly identified.

Table 1. The parameters of simulated datasets.

Data ID	Sample size	MAFs	HP
Pure1	1000	(0.2, 0.2)	1.0
Pure2	2000	(0.2, 0.2)	1.0
Pure3	3000	(0.2, 0.2)	1.0
Pure4	4000	(0.2, 0.2)	1.0
Pure5	8000	(0.2, 0.2)	1.0
Pure6	1000	(0.2,0.2,0.2)	1.0
Pure7	2000	(0.2,0.2,0.2)	1.0
Pure8	3000	(0.2,0.2,0.2)	1.0
Pure9	4000	(0.2,0.2,0.2)	1.0
Pure10	8000	(0.2,0.2,0.2)	1.0
Hete1	1000	(0.2, 0.2) (0.3,0.3)	H1 = 50%, H2 = 50%
Hete2	2000	(0.2, 0.2) (0.3,0.3)	H1 = 50%, H2 = 50%
Hete3	3000	(0.2, 0.2) (0.3,0.3)	H1 = 50%, H2 = 50%
Hete4	4000	(0.2, 0.2) (0.3,0.3)	H1 = 50%, H2 = 50%
Hete5	8000	(0.2, 0.2) (0.3,0.3)	H1 = 50%, H2 = 50%
Hete6	1000	(0.2,0.2,0.2) (0.3,0.3,0.3)	H1 = 50%, H2 = 50%
Hete7	2000	(0.2,0.2,0.2) (0.3,0.3,0.3)	H1 = 50%, H2 = 50%
Hete8	3000	(0.2,0.2,0.2) (0.3,0.3,0.3)	H1 = 50%, H2 = 50%
Hete9	4000	(0.2,0.2,0.2) (0.3,0.3,0.3)	H1 = 50%, H2 = 50%
Hete10	8000	(0.2,0.2,0.2) (0.3,0.3,0.3)	H1 = 50%, H2 = 50%

4. Experimental results

In this section, our method MCMC (maximum correlation and maximum consistence) was compared with DPEH [34] and MDR [57]. The DPEH is a three-stage framework for epistasis detection, heterogeneity analysis and disease prediction based on deep learning. Note that DPEH adopted a ESMO method which applied an exhaustive search strategy to search epistasis with using two optimization objectives. The MDR reduces high dimensional multi-locus genotype into one dimension and also uses an exhaustive search strategy to find the riskiest epistatic combinations.

4.1. Results on pure datasets

Since a sample in pure dataset only has two possible states: normal or case, there is no need to distinguish tumor subtypes, so that the diagnosis can be regarded as a binary classification problem.

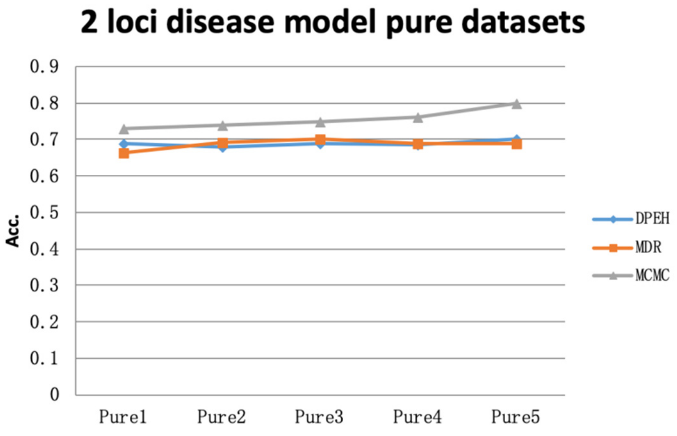

Figure 2. The accuracy results of pure datasets simulated by 2 loci disease model.

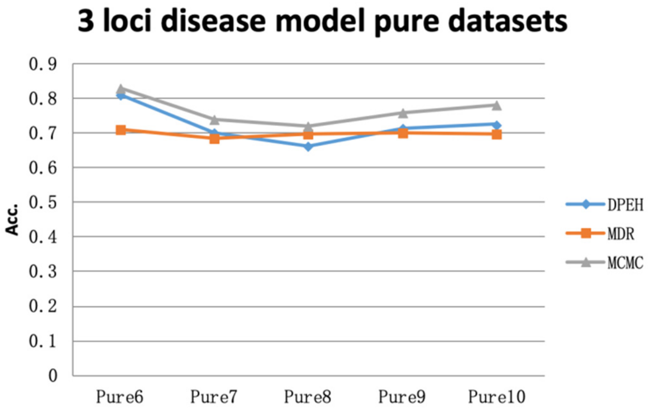

Figure 3. The accuracy results of pure datasets simulated by 3 loci disease model.

It can be seen from the results in the Figures 2 and 3 that the accuracy of our method MCMC is significantly improved compared with the other two methods, and the accuracy is increased by more than 5%. The increase in accuracy mainly comes from XGBoost's ability to more accurately describe the relationship between epistasis and disease status. In addition, on average, the prediction accuracy of the 3 loci dataset is slightly higher than that of the 2 loci dataset. The reason for this phenomenon may be that the increase in the number of features input to the prediction model XGBoost provides more information.

4.2. Results on heterogeneous datasets

The heterogeneous simulation data set is generated by two kinds of pathogenic models, so there are 2 subtypes in the cases, and there are 3 categories in addition to the normal sample, so it is a triple classification problem. Since the MDR cannot directly deal with heterogeneous data, we use the same processing method as the DPEH to compare the accuracy of prediction with the MDR.

4.2.1. The accuracy of 2 loci disease models

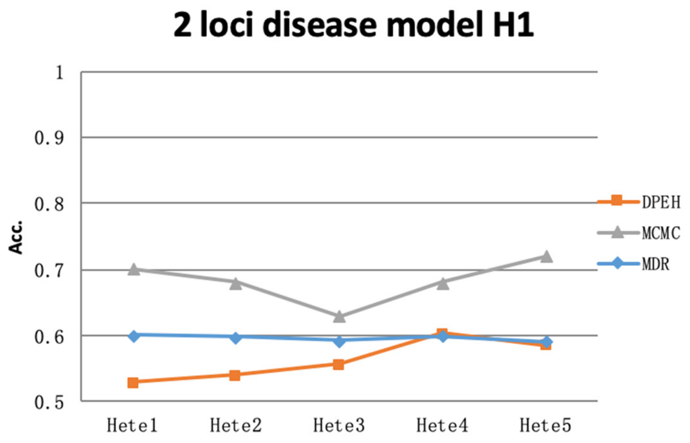

Figure 4. The accuracy results of heterogeneous datasets simulated by 2 loci disease model H1.

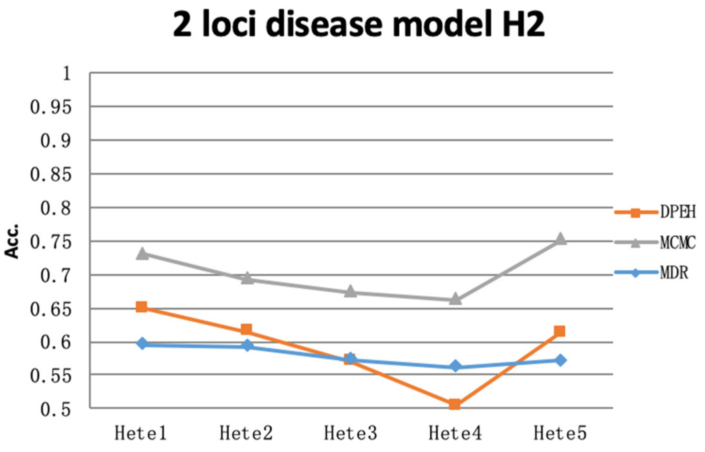

Figure 5. The accuracy results of heterogeneous datasets simulated by 2 loci disease model H2.

From the results in the Figures 4–6, it can be found that the MCMC is significantly improved compared with the subtype prediction of the DPEH and MDR methods. And it can be found that the accuracy of the pathogenic loci corresponding to the H2 model shown in Figure 5 as predictive features is higher than that of the H1 model shown in Figure 4. The reason may be that the MAFs of the H2 model are higher than the H1. It means that this mutation pattern of H2 in the case is more stable H1, which facilitates subtype identification. Comparing Figure 6 with Figures 4 and 5, it can be found that jointly inputting the pathogenic loci of the two models into the prediction model at the same time can further improve the prediction accuracy, because the information inputted to the model is more complete.

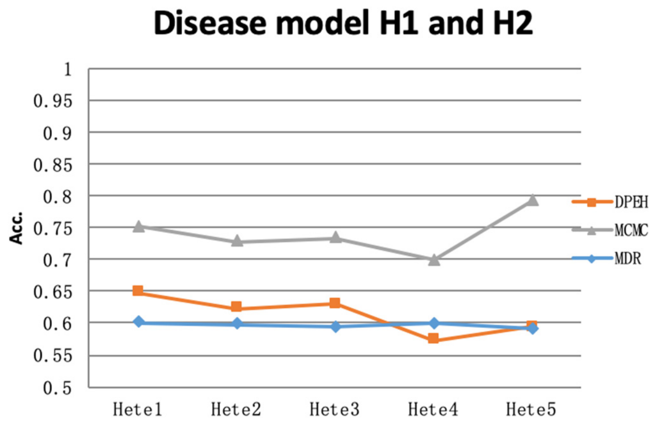

Figure 6. The accuracy results of heterogeneous datasets simulated by 2 loci disease models H1 and H2.

4.2.2. The accuracy of 3 loci disease models

In Figures 7–9, we compared the results of DPEH, MDR and MCMC on 3 loci disease models. It can be found from the experimental results that the performance of the MCMC method is still significantly higher than the DPEH and MDR methods in this disease model. Similarly, the susceptibility loci corresponding to the H2 disease model can train more accurate subtype classifiers. In addition, the joint input of the pathogenic loci of the two disease models into the classifier can further improve the accuracy.

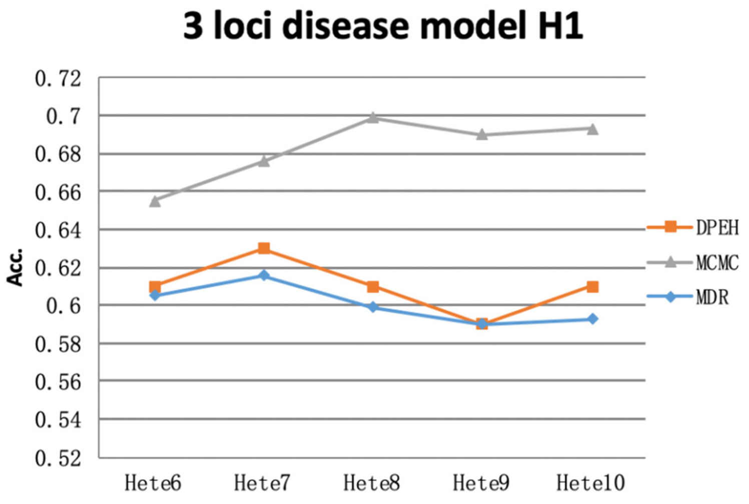

Figure 7. The accuracy results of heterogeneous datasets simulated by 3 loci disease model H1.

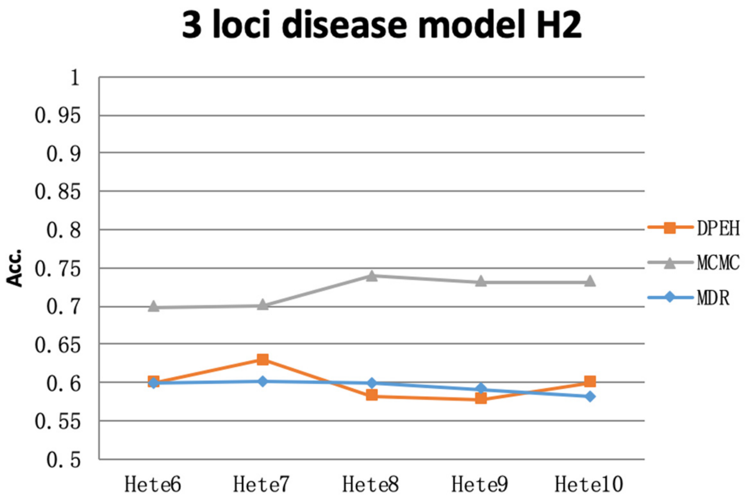

Figure 8. The accuracy results of heterogeneous datasets simulated by 3 loci disease model H2.

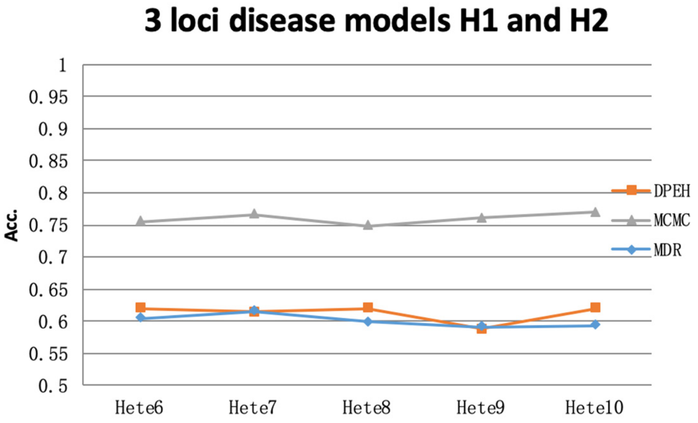

Figure 9. The accuracy results of heterogeneous datasets simulated by 3 loci disease models H1 and H2.

Table 2. The results of DPEH and MCMC on power measure.

	DPEH		MCMC	
Data ID	Power of H1 (%)	Power of H2 (%)	Power of H1 (%)	Power of H2 (%)
Hete1	68	72	68	**75**
Hete2	65	68	**69**	**70**
Hete3	65	65	63	65
Hete4	72	75	**75**	**78**
Hete5	62	62	**65**	**70**
Hete6	56	58	**60**	**62**
Hete7	55	55	**58**	**60**
Hete8	58	60	**60**	**62**
Hete9	63	65	**68**	**65**
Hete10	60	63	**66**	**68**

4.2.3. The power on heterogeneous datasets

Table 2 details the detection efficiency of the two methods for each disease model. The bold numbers in the experimental results indicate that the performance of MCMC is more advantageous, so it can be found that in most cases, the performance of the MCMC method is significantly better than the DPEH method. More importantly, disease models with larger MAF are easier to accurately identify, and the detection efficiency of the 2 loci disease model is better than that of the 3 loci disease model.

5. Discussion and conclusions

Tumor epistasis makes it difficult to find weakly effect susceptibility genes, while tumor heterogeneity makes it difficult to distinguish patient subtypes. Solving these two problems is the basic work of tumor precision medicine. In practical application, for tumors whose subtypes have not been fully defined, we can first treat the cohort for genome sequencing, and then use our method MCMC to analyze epistasis and heterogeneity, so as to fully identify tumor subtypes with different combinations of pathogenic sites. On the basis of fully defining tumor subtypes, our method can be used to classify new samples. Our computational framework proposed in this study has the following three advantages:

1) The MCMC method has more significant advantages than previous methods in terms of efficiency and accuracy.

2) The framework first uses improved genetic algorithms to optimize different epistasis evaluation criteria to identify multiple potential epistasis risk combinations. This multi-objective evaluation criterion improves the possibility of identifying weakly effect susceptible genes from different perspectives. Therefore, it is more conducive to the comprehensive discovery of target genes of complex diseases.

3) Adaptively determine the number of tumor subtypes based on high-confidence risk combinations, and then use a clustering algorithm to identify heterogeneity in case samples. Therefore, the MCMC method can adaptively deal with epistasis and heterogeneity at the same time, and has certain practical value.

Although the MCMC method designed in this study has achieved certain advantages in evaluation measures, there are still areas for further improvement. For example, there is still room for improvement in the efficiency of epistatic recognition and the classification accuracy of tumor subtypes. More important, it is necessary to apply our method to real tumor case-control study to further verify the reliability of tumor heterogeneity from the perspectives of tumor tissue morphology and tumor immune escape pathways.

Acknowledgments

This work was partially supported by Natural Science Foundation of Hunan Province of China (NO. 2020JJ7081). And, it was also supported by the Jiangxi Provincial natural science fund (No. 20192ACB21004 and 20204BCJL23035).

Conflict of interests

All authors declare no conflicts of interest in this paper.

References

1. E. A. Ashley, Towards precision medicine, *Nat. Rev. Genet.*, **17** (2016), 507.
2. H. Peng, X. Zeng, Y. Zhou, D. Zhang , R. Nussinov, F. Cheng, A component overlapping attribute clustering (COAC) algorithm for single-cell RNA sequencing data analysis and potential pathobiological implications, *PLoS Comput. Biol.*, **15** (2019), e1006772.
3. X. Liu, Z. Hong, J. Liu, Y. Lin, R. Alfonso, Q. Zou, et al, Computational methods for identifying the critical nodes in biological networks, *Briefings Bioinf.*, **21** (2020), 486–497.
4. A. Alizadeh, V. Aranda, A. Bardelli, C. Blanpain, C. Bock, C. Borowski, et al., Toward understanding and exploiting tumor heterogeneity, *Nat. Med.*, **21** (2015), 846–853
5. Q. Jia, W. Wu, Y. Wang, P. B. Alexander, C. Sun, Z. Gong, et al, Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer, *Nat. Commun.*, **9** (2018), 1–10.
6. Y. Gu, Y. Gao, X. Tang, H. Xian, K. Shi, Bioinformatics analysis identifies CPZ as a tumor immunology biomarker for gastric cancer, *Curr. Bioinf.*, **16** (2021), 98–105.
7. W. Ran, X. Chen, B. Wang, Y. Ping, X. Xiao, Whole-exome sequencing of tumor-only samples reveals the association between somatic alterations and clinical features in pancreatic cancer, *Curr. Bioinf.*, **15** (2020), 1160–1167.
8. Z. Lv, F. Cui, Q. Zou, L. Zhang, L. Xu, Anticancer peptides prediction with deep representation learning features, *Briefings Bioinf.*, (2021), bbab008.
9. A. C. Iliopoulos, G. Beis, P. Apostolou, I. Papasotiriou, Complex networks, gene expression and cancer complexity: a brief review of methodology and applications, *Curr. Bioinf.*, **15** (2020), 629–655.
10. A. Ghosh, H. Yan, Stability analysis at key positions of EGFR related to non-small cell lung cancer, *Curr. Bioinf.*, **15** (2020), 260–267.
11. Z. Ramzan, M. A. Hassan, H. M. S. Asif, A. Farooq, A Machine Learning-based Self-risk Assessment Technique for Cervical Cancer, *Curr. Bioinf.*, **16** (2021), 315–332.
12. Y. Luo, X. Wang, L. Li, Q. Wang and Y. Luo, Bioinformatics analysis reveals centromere protein K can serve as potential prognostic biomarker and therapeutic target for non-small cell lung cancer, *Curr. Bioinf.*, **16** (2021), 106–119.
13. S. Liu, H. Tang, H. Liu, J. Wang, Multi-abel learning for the diagnosis of cancer and identification of novel biomarkers with high-throughput omics, *Curr. Bioinf.*, **16** (2021), 261–273.
14. L. Yang, H. Gao, K. Wu, H. Zhang, L. Tang, Identification of cancerlectins by using cascade linear discriminant analysis and optimal g-gap tripeptide composition, *Curr. Bioinf.*, **15** (2020), 528–537.
15. Z. Lv, J. Zhang, H. Ding, Q. Zou, RF-PseU: a random forest predictor for RNA pseudouridine sites, *Front. Bioeng. Biotechnol.*, **8** (2020), 134.
16. Q. Yang, B. Li, S. Chen, T. Jing, Y. Li, L. Yi, et al, MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis, *J. Proteomics*, (2021), 104023.
17. Z. Lv, P. Wang, Q. Zou, Q. Jiang, Identification of Sub-Golgi protein localization by use of deep representation learning features, *Bioinformatics*, **36** (2020), 5600–5609.
18. F. Wang, G. Qin, J. Liu, X. Wang, B. Ye, Bio-analytical identification of key genes that could contribute to the progression and metastasis of osteosarcoma, *Curr. Bioinf.*, **16** (2021), 216–224.

19. Z. Lv, S. Jin, H. Ding, Q. Zou, A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features, *Front. Bioeng. Biotechnol.*, **7** (2019).
20. Z. Lv, C. Ao, Q. Zou, Protein function prediction: from traditional classifier to deep learning, *Proteomics*, **19** (2019), 1900119.
21. J. Pan, X. Luo, T. Shao, C. Li, G. Wang, Identification of genomic islands in synechococcus sp. WH8102 using genomic barcode and whole-genome microarray analysis, *Curr. Bioinf.*, **16** (2021), 24–30.
22. H. Wang, Y. Ding, J. Tang, F. Guo, Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt Independence Criterion, *Neurocomputing*, **383** (2020), 257–269.
23. Y. Shen, J. Tang, F. Guo, Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC, *J. Theor. Biol.*, **462** (2019), 230–239.
24. Y. Ding, J. Jun, G. Fei, Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion, *Knowl.-Based Syst.*, **204** (2020), 106254.
25. H. Wang, Y. Ding, J. Jun, G. Fei, Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment, *Briefings Bioinf.*, 2021.
26. Y. Ding, J. Tang, F. Guo, Identification of drug-target interactions via fuzzy bipartite local model, *Neural. Comput. Appl.*, **23** (2020), 10303–10319.
27. X. Zeng, X. Song, T. Ma, X. Pan, Y. Zhou, Y Hou, Repurpose open data to discover therapeutics for COVID-19 using deep learning, *J. Proteome Res.*, **19** (2020), 4624–4636.
28. G. Turashvili, E. Brogi, Tumor heterogeneity in breast cancer, *Front. Biomed.*, (2017), 227.
29. M. Hofree, J. P. Shen, H. Carter, G. Andrew, I. Trey, Network-based stratification of tumor mutations, *Nat. Med.*, **10** (2013), 1108–1115.
30. X. Li, A fast and exhaustive method for heterogeneity and epistasis analysis based on multi-objective optimization, *Bioinformatics*, **33** (2017), 2829–2836.
31. H. Xu, W. Zeng, D. Zhang, X. Zeng, MOEA/HD: a multiobjective evolutionary algorithm based on hierarchical decomposition, *IEEE. Trans. Cybern.*, **49** (2019), 517–526.
32. H. Xu, W. Zeng, X. Zeng, G. Gary, An evolutionary algorithm based on Minkowski distance for many–objective optimization, *IEEE. Trans. Cybern.*, **49** (2019), 3968–3979.
33. X. Zeng, W. Wang, C. Chen, G. Yen, A consensus community-based particle swarm optimization for dynamic community detection, *IEEE. Trans. Cybern.*, **50** (2020), 2502–2513.
34. X. Li, C. Wang, L. Liu, X. Xia, A method for heterogeneity analysis of complex diseases based on clustering algorithm, in *2017 13th International Conference on Computational Intelligence and Security (CIS)*, (2017).
35. L. Jiang, Y. Ding, J. Tang, G. Fei, MDA-SKF: similarity kernel fusion for accurately discovering miRNA-disease association, *Front. Genet.*, **9** (2018), 1–13.
36. B. Liu, X. Gao, H. Zhang, BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches, *Nucleic Acids Res.*,**47** (2019), e127.
37. B. Liu, C. Li, K. Yan, DeepSVM-fold: Protein fold recognition by combining Support Vector Machines and pairwise sequence similarity scores generated by deep learning networks, *Briefings Bioinf.*, **21** (2020), 1733–1741.

38. S. Jin, X. Zeng, F. Xia, W. Huang, X. Liu, Application of deep learning methods in biological networks, *Briefings Bioinf.*, **22** (2021), 1902–1917.
39. L. Cai, L. Wang, X. Fu, C. Xia, X. Zeng, Q. Zou, ITP-Pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation, *Briefings Bioinf.*, **22** (2020).
40. H. Lv, F. Dao, Z. Guan, H. Yang, Y. Li, H. Lin, Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method, *Briefings Bioinf.*, **22** (2021).
41. F. Dao, H. Lv, D. Zhang, Z. Zhang, H. Lin, DeepYY1: a deep learning approach to identify YY1-mediated chromatin loops, *Briefings Bioinf.*, **22** (2021).
42. F. Dao, H. Lv, W. Su, Z. Sun, H. Lin, iDHS-Deep: an integrated tool for predicting DNase I hypersensitive sites by deep neural network, *Briefings Bioinf.*, 2021.
43. D. Wang, Z. Zhang, Y. Jiang, Z. Mao, D. Xu, DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism, *Nucleic Acids Res.*, **49** (2021), e46–e46.
44. Y. Jiang, D. Ma, C. Suo, Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies, *Cancer Cell*, **35** (2019), 428–440.
45. M. J. Bou-Dargham, Y. Liu, Q. Sang, J. Zhang, T. Seagroves, Subgrouping breast cancer patients based on immune evasion mechanisms unravels a high involvement of transforming growth factor-beta and decoy receptor 3, *PLoS One.*, **13** (2018), e0207799.
46. A. Robertson, J. Shih, C. Yau, E. Gibb, J. Oba, K. Mungall, et al, Integrative analysis identifies four molecular and clinical subsets in uveal melanoma, *Cancer Cell*, **32** (2017), 204–220.
47. Z. Xiong, Q. Yang, X. Li, Effect of intra-and inter-tumoral heterogeneity on molecular characteristics of primary IDH-wild type glioblastoma revealed by single-cell analysis, *CNS. Neurosci. Ther.*, **26** (2020), 981–989.
48. D. Lawson, K. Kessenbrock, R. Davis, N. Pervolarakis, Z. Werb, Tumour heterogeneity and metastasis at single-cell resolution, *Nat. Cell Biol.*, **20** (2018), 1349–1360.
49. H. Xu, W. Zeng, X. Zeng, G. Yen, A polar-metric-based evolutionary algorithm, *IEEE. Trans. Cybern.*, 2021.
50. P. Jing, H. Shen, MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies, *Bioinformatics*, **31** (2015), 634–641.
51. M. Srinivas, L. Patnaik, Adaptive probabilities of crossover and mutation in genetic algorithms, *IEEE. Trans. Syst. Man. Cybern.* **24** (1994), 656–667.
52. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, (2016), 785–794.
53. R. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. H. Moore, GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures, *BioData Min.*, **5** (2012), 1–14.
54. L. Wei, Y. Ding, R. Su, Y. Jie, S. Ran, Prediction of human protein subcellular localization using deep learning, *J. Parallel. Distr. Com.*, **117** (2018), 212–217.
55. L. Wei, W. He, A. Malik, B. Manavalan, Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework, *Briefings Bioinf.*, **22** (2021).
56. Y. Ding, J. Tang, F. Guo, Identification of drug-side effect association via multiple information integration with centered kernel alignment, *Neurocomputing*, **325** (2019),211–224.

57. J. Moore, J. Gilbert, C. Tsai, F. Chiang, T. Holden, N. Barney, et al, A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility, *J. Theor. Biol.*, **241** (2006), 252–261.