Briefings in Bioinformatics, 22(6), 2021, 1–11

https://doi.org/10.1093/bib/bbab100
Case Study

Integration and interplay of machine learning and bioinformatics approach to identify genetic interaction related to ovarian cancer chemoresistance

Kexin Chen†, Haoming Xu†, Yiming Lei, Pietro Lio, Yuan Li, Hongyan Guo and Mohammad Ali Moni

Corresponding authors: Mohammad Ali Moni, WHO Collaborating Centre on eHealth, UNSW Digital Health, School of Public Health and Community Medicine, Faculty of Medicine, University of New South Wales, Sydney, NSW 2052, Australia. Telephone: +61414701759, E-mail: m.moni@unsw.edu.au; Yiming Lei, Block 2, Science Building, School of EECS, Peking University, Haidian District, Beijing, China. Telephone: 86-10-62751774, E-mail: leiym@pku.edu.cn; Hongyan Guo, Department of Obstetrics and Gynecology, Peking University Third Hospital, Haidian District, Beijing, China. Telephone: +86-10-82267510, E-mail: bysyghy@163.com.
†Kexin Chen and Haoming Xu contributed equally to this work.

Abstract

Although chemotherapy is the first-line treatment for ovarian cancer (OCa) patients, chemoresistance (CR) decreases their progression-free survival. This paper investigates the genetic interaction (GI) related to OCa-CR. To decrease the complexity of establishing gene networks, individual signature genes related to OCa-CR are identified using a gradient boosting decision tree algorithm. Additionally, the genetic interaction coefficient (GIC) is proposed to measure the correlation of two signature genes quantitatively and explain their joint influence on OCa-CR. Gene pair that possesses high GIC is identified as signature pair. A total of 24 signature gene pairs are selected that include 10 individual signature genes and the influence of signature gene pairs on OCa-CR is explored. Finally, a signature gene pair-based prediction of OCa-CR is identified. The area under curve (AUC) is a widely used performance measure for machine learning prediction. The AUC of signature gene pair reaches 0.9658, whereas the AUC of individual signature gene-based prediction is 0.6823 only. The identified signature gene pairs not only build an efficient GI network of OCa-CR but also provide an interesting way for OCa-CR prediction. This improvement shows that our proposed method is a useful tool to investigate GI related to OCa-CR.

Key words: chemoresistance; ovarian cancer; gene pair; genetic interaction

Kexin Chen got her BSc degree from Yuanpei College, Peking University, China in 2020. Now she is an external research assistant at School of Electronics Engineering and Computer Science, Peking University, China. Her research interest includes medical imaging and bioinformatics.
Haoming Xu is visiting scholar in Department of Biomedical Engineering, Duke University, USA, and assistant professor at Modern Science and Technology Laboratory, Chengdu Institute of Public Administration, China. He received his MPhil in Advanced Computer Science from the University of Cambridge.
Yiming Lei is a research assistant professor at School of Electronics Engineering and Computer Science, Peking University, China, and assistant director at Engineering Research Center of Mobile Digital Hospital Systems, Ministry of Education, China. His research interest includes medical imaging, biophysics and bioinformatics.
Pietro Liò is a full professor of Computational Biology and a member of the Artificial Intelligence Group at the University of Cambridge. His research interest focuses on bioinformatics and machine learning related to biological complexity.
Yuan Li is currently working as an attending doctor at Department of Gynecology, Peking University Third Hospital, China. Her research interests include gynecology oncology, endometriosis and genetic consulting.
Hongyan Guo is the head of Department of Gynecology, Peking University Third Hospital, China. Also, she is the head of Department of Gynecology, Health Science Center, Peking University, China. She has authored more than 30 peer reviewed papers. She presided over National Key Technologies R&D Program and other provincial and ministerial level projects. Her research interests include gynecology oncology, endometriosis, and minimally invasive surgery.
Mohammad Ali Moni is a senior research fellow and conjoint lecturer at the University of New South Wales, Australia. He received his PhD in Artificial intelligence and Clinical bioinformatics from the University of Cambridge. His research interest encompasses artificial intelligence, machine learning, data science, health informatics and clinical bioinformatics.
Submitted: 13 January 2021; **Received (in revised form):** 4 March 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

Introduction

Ovarian cancer (OCa) is a common type of cancer that has a high mortality rate. Chemotherapy is the first-line treatment for OCa patients, but chemoresistance (CR) decreases their progression-free survival (PFS). Now, it is accepted that the cancer-like diseases are affected by individual genes, and the latter also shows the influence on CR [1, 2]. Machine learning has been used to identify the CR-related signature genes, which helps the researchers to select proper therapeutics and predict drug responses [3–5]. Recent studies reveal that cancers and CR are usually affected by both individual genes and their interactions [6, 7]. A reason that is often cited for the lack of success in genetic studies of complex disease is the existence of interactions between individual signature genes. If an individual signature gene functions primarily through a complex mechanism that involves multiple other genes, the effect might be missed if the gene is examined in isolation without allowing for its potential interactions with other genes [8, 9].

The genetic interaction (GI) indicates that the effect of one gene is related to that of another gene that helps the researchers to delineate the pathways, protein complexes and underlying biological processes [10–12]. Machine learning models are widely used to identify the relationships between features and discover GIs [13, 14]. The previous study red[15] identified hundreds of signature gene pairs simultaneously and used gene pair-based machine learning model to make prediction. These interactions are important for delineating functional relationships among genes and their corresponding proteins [16].

This paper presents our efforts to analyze the GI related to OCa-CR. Considering that GI network in previous research is complicated due to its high dimension, a small number of individual signature genes are desired. Aiming to identify the individual signature genes, we constructed an ensemble learning model based on gradient boosting decision tree (GBDT) [17]. The GBDT is a widely used machine learning algorithm, which obtains state-of-the-art results on many machine learning tasks, and has been actively used in computational biology and bioinformatics [18, 19]. To build an efficient small-size GI network, the genetic interaction coefficient (GIC) method is proposed to measure the correlations among individual signature genes. In the following, the influence of signature gene pairs on OCa-CR is discussed. Finally, it is shown that these signature gene pairs show satisfactory performance in predicting OCa-CR.

Table 1 presents the list of the abbreviations and their full names used throughout the paper.

Table 1. List of abbreviations used throughout the paper

Abbreviation	Full Name
OCa	Ovarian cancer
CR	Chemoresistance
PFS	Progression-free survival
GI	Genetic interaction
GBDT	Gradient boosting decision tree
GIC	Genetic interaction coefficient
AUC	Area under curve
TCGA	The cancer genome atlas
SMOTE	Synthetic minority oversampling technique
CART	Classification and regression tree
TP	True positive
FP	False positive
TN	True negative
FN	False negative
TG	Target group
SG	Standard group
GO	Gene ontology

Materials and methods

Overview of analytical approach

We presented here, summarized in Figure 1, the workflow of this research that consists of the identification of individual signature genes, identification of signature gene pairs and prediction based on signature gene pairs.

Data preprocessing

The cancer genome atlas (TCGA) database [20] provides the data sets of totally 35 OCa-CR and 162 OCa-CS patients. The data set of each TCGA sample shows the expression levels of 14252 individual genes and includes PFS data (stating from initial treatment) as well. The samples whose PFS are smaller than 9 months are classified into OCa-CR group, and those whose PFS are higher than 15 months are classified into the OCa-CS group. From clinical treatment perspective, the patient is considered to be OCa-CR when the recurrence time is shorter than 6 months. This recurrence time is counted from the ends of treatment. However, the PFS data from TCGA are counted from the start of the treatment and clinical treatment usually takes 6 months. From this definition, 12 months of PFS can divide the patients into OCa-CS and OCa-CR group. In order to avoid noise and classification bias, we choose 9 months and 15 months to divide OCa-CS and OCa-CR group. Genes that have expression levels of zero are eliminated from data sets.

Data preprocessing includes data normalization and data oversampling. In normalization of the gene expressions of each patient, the method of Exploratory Data Analysis and Normalization for Sequence data (EDASeq) is applied. EDASeq method includes two normalization steps: a within-sample normalization step that adjusts for gene-specific and sample-specific effects and a between-sample normalization that corrects distributional differences between samples [21]. Besides, gene data from TCGA usually has the problem of class imbalance, which may affect model training and its subsequent performance [22]. To balance the OCa-CR and OCa-CS samples, OCa-CR patients are oversampled. There are many different types of data oversampling method for a typical classification problem. The most common technique is called synthetic minority oversampling technique (SMOTE) algorithm, which oversamples the minority class by creating synthetic minority class examples [23]. Here SMOTE algorithm is applied in the oversampling of OCa-CR samples [24].

After the data preprocessing, totally 324 samples (162 OCa-CR samples and 162 OCa-CS samples) are finally obtained. When dividing the data into a training set and a test set, we use hold-out method, a widely used method in bioinformatics and gene research [25], which randomly selects 30% of the data as test set. Thus, the training set includes 226 samples and the test set includes 98 samples.

GBDT modeling

Ensemble learning, includes GBDT and AdaBoost, is a widely used methodology in genetic research. Among different ensemble learning methods, GBDT possesses outstanding predicting performance in the analysis of high dimensional gene data [26, 27]. The GBDT is a type of ensemble model that consists of

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

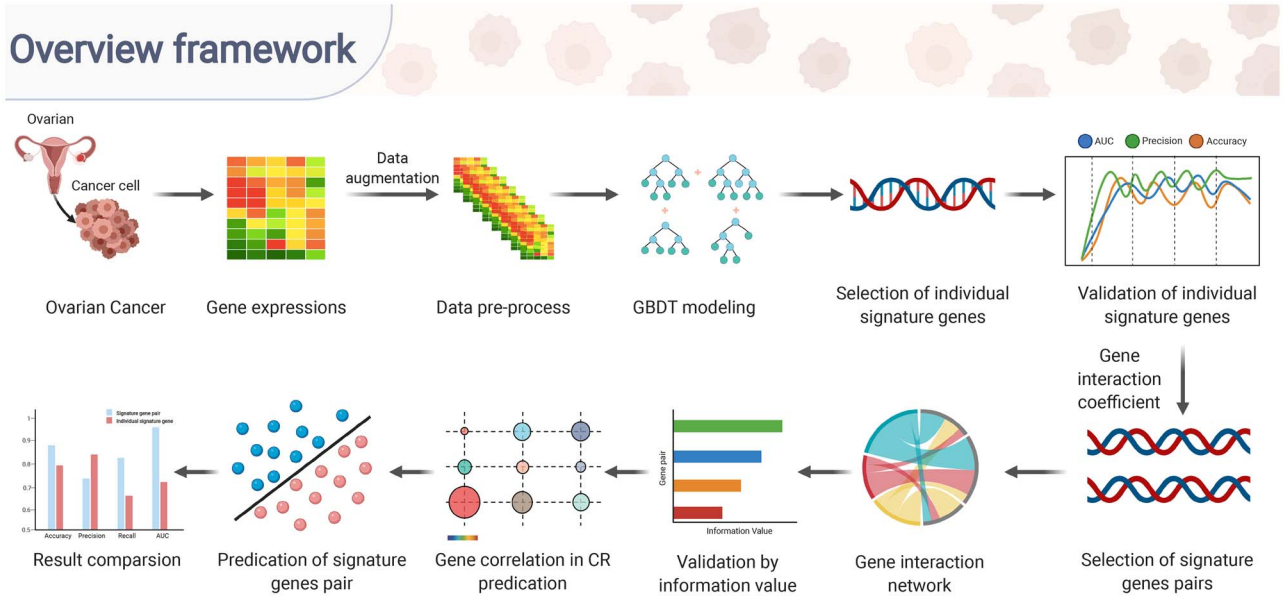

Figure 1. Workflow of this research that includes the identification of individual signature genes, identification of signature gene pairs and prediction based on signature gene pairs.

a series of decision trees. In GBDT modeling, gradient boosting promotes the performance gradually by reducing the residual. In each iteration, the model is refreshed to fit the negative gradient of the loss function until it converges. The final prediction is the summation of former model results. In GBDT modeling, classification and regression trees (CARTa) decision tree is embedded into gradient boosting as the basic weak learner in each iteration [28].

Training set $\{(x_1, y_1), \ldots (x_n, y_n)\}$ consists of $n = 226$ samples. Here x_i represents the gene expression level of patient i, whereas y_i indicates whether patient i belongs to OCa-CR group. $F(x)$ refers to the functional relationship between gene expressions and OCa-CR. Using the training set, the goal of the algorithm is to find out an approximation $\hat{F}(x)$ to the function $F(x)$ that minimizes the expected value of some specified loss function $L(y, F(x))$ and identify individual signature genes accordingly:

$$\hat{F} = \arg\min_{F} E_{x,y}\left[L(y, F(x))\right] \tag{1}$$

In order to improve predicting accuracy of identified individual signature genes, gradient boosting seeks an approximation in the form of a weighted sum of weak learners $h_1(x), h_2(x), \ldots, h_M(x)$, i.e.

$$\hat{F} = \sum_{m=1}^{M} \rho_m h_m(x) \tag{2}$$

Here we set $M = 100$ to avoid over-fitting and under-fitting. In the iteration process, GBDT solving the binary classification problem has the form

$$F_0(x) = 0.5 * \log\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1 - y_i)}\right) \tag{3}$$

and

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x) \tag{4}$$

The ensemble model uses $\{(x_1, r_{1m}) \cdots (x_n, r_{nm})\}$ to train a Weak learner decision tree $h_m(x)$, where r_{im} is the pseudo-residuals, i.e.

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{5}$$

Here, the parameter ρ_m is selected to ensure that the gradient of $F_m(x)$ is the one that makes the loss function of the former model $F_{m-1}(x)$ decreases fastest, therefore ρ_m has the form

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \rho h_m(x_i)) \tag{6}$$

Then, the individual signature genes are selected according to their contribution to the classification process of GBDT and its weak learners. In GBDT modeling, decision tree (especially CART tree) is regarded as weak learner $h_m(x)$, which has the form

$$h_m(x) = \sum_{j=1}^{J} b_{jm} I(x \in R_{jm}) \tag{7}$$

where J is the number of its leaves. The tree partitions the input space into J disjoint regions $R_{1m}, \ldots, R_{Jm}$ and predicts a constant value in each region. b_{jm} is the value predicted in R_{jm}.

Feature importance quantifies each gene's contribution to the model. Thus, individual signature genes can be selected to acquire better interpretability based on the feature importance of GBDT modeling.

Feature importance of GBDT is calculated by the average feature importance over all of the CART trees in the model. CART tree uses Gini impurity criterion to split the node and create decision tree. Based on Gini impurity, the feature importance of variable j in tree T can be calculated as

$$FI_j(T) = \sum_{t=1}^{L} \Delta_{\text{Gini}} I(v_t = j) \tag{8}$$

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

which is the summation of the non-terminal nodes t of the tree T, v_t is the splitting variable associated with node t, and Δ_{Gini} is the corresponding decrease of Gini impurity. For a series of CART trees $\{T_m\}$ obtained through gradient boosting approach, feature importance of variable j can be generalized by the average value of all the trees in the sequence

$$\text{FI}_j = \frac{1}{M}\sum_{m=1}^{M}\text{FI}_j(T_m) \tag{9}$$

In order to improve the stability of the algorithm and avoid the influence of randomness, the experiment is repeated 20 times. Let $\text{FI}_j^{(k)}$ denotes the feature importance of variable j in kth experiment. $\text{FI}_j^{(k)} > 0$ represents that variable j are used in kth experiment, whereas $\text{FI}_j^{(k)} = 0$ means that variable j does not contribute to the model. Thus the frequency of variable j in our experiments, f_j, can be defined as

$$f_j = \sum_{k=1}^{20} sign(\text{FI}_j^{(k)}) \tag{10}$$

where $sign(\cdot)$ is the sign function that has the form:

$$sign(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \end{cases} \tag{11}$$

The identified individual signature genes are those that make a stable contribution to the model. This contribution is quantified by frequency f_j in Equation (10). The threshold is manually set as half of the experiment times N. In our experiments, we set experiment times $N = 20$.

Analysis method

This section validates the identified individual signature genes. In the validation using data processing indicators, the accuracy, precision and area under curve (AUC) values are calculated. The identified individual signature genes are also validated using published biological literature. For properly evaluating the identified individual signature genes, we established a random forest classifier to calculate above data processing indicators [29]. Their definitions are listed as follows.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{14}$$

For each patient in the test set, if the prediction outcome is CR and the actual situation is also CR, then it is called a true positive (TP); however, if the actual situation is CS, then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction and the actual situation are CS, and false negative (FN) is when the prediction is CS, whereas the actual situation is CR.

The area under the Receiver Operating Characteristic (ROC) curve is also used in the validation process [30]. ROC is created by plotting the recall against the specificity at various threshold settings. Thus AUC is calculated as

$$\text{AUC} = \int_{x=0}^{1}\text{Recall}(\text{Specificity}^{-1}(x))dx \tag{15}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{16}$$

Genetic interaction coefficient

In machine learning, the research on feature correlation attracts clear attentions [31]. GIC is proposed to measure the correlation of two variables quantitatively and explain their joint influence on the dependent variable. To calculate GIC of each gene pair, the OCa samples are divided into two groups, target group (TG) and standard group (SG), according to the dependent variable α. We assume $\beta = 1$ in TG, whereas $\beta = 0$ in SG. The calculation of GIC includes three steps: normalization, discretization and Hadamard product. ${}^{\alpha}_{\beta}a_{i,j}\big|_{\beta=1}$ is gene j's expression level of patient i in TG. Firstly, ${}^{\alpha}_{\beta}b_{i,j}\big|_{\beta=1}$ is the result after applying normalization on ${}^{\alpha}_{\beta}a_{i,j}\big|_{\beta=1}$, and n_α is the number of samples in TG.

$${}^{\alpha}_{\beta}b_{i,j}\big|_{\beta=1} = \frac{{}^{\alpha}_{\beta}a_{i,j}\big|_{\beta=1} - {}^{\alpha}_{\beta}\mu_j\big|_{\beta=0}}{{}^{\alpha}_{\beta}\sigma_j\big|_{\beta=0}} \tag{17}$$

where ${}^{\alpha}_{\beta}\mu_j\big|_{\beta=0}$ is the mean of gene j's expression level in SG, i.e.

$${}^{\alpha}_{\beta}\mu_j\big|_{\beta=0} = \frac{\sum_{i=1}^{n_\alpha}{}^{\alpha}_{\beta}a_{i,j}\big|_{\beta=0}}{n_\alpha} \tag{18}$$

where ${}^{\alpha}_{\beta}\sigma_j\big|_{\beta=0}$ is the standard deviation of gene j's expression level in SG, i.e.

$${}^{\alpha}_{\beta}\sigma_j\big|_{\beta=0} = \sqrt{\frac{\sum_{i=1}^{n_\alpha}\left({}^{\alpha}_{\beta}a_{i,j}\big|_{\beta=0} - {}^{\alpha}_{\beta}\mu_{i,j}\big|_{\beta=0}\right)^2}{n_\alpha}} \tag{19}$$

Secondly, ${}^{\alpha}_{\beta}b_{i,j}\big|_{\beta=1}$ is discretized into ${}^{\alpha}_{\beta}c_{i,j}\big|_{\beta=1}$ as follows:

$${}^{\alpha}_{\beta}c_{i,j}\big|_{\beta=1} = \begin{cases} 1 & {}^{\alpha}_{\beta}b_{i,j}\big|_{\beta=1} > \lambda_1 \\ 0 & \lambda_2 \le {}^{\alpha}_{\beta}b_{i,j}\big|_{\beta=1} \le \lambda_1 \\ -1 & {}^{\alpha}_{\beta}b_{i,j}\big|_{\beta=1} < \lambda_2 \end{cases} \tag{20}$$

For the convenience of following clarifications, here the upper bar and lower bar of ${}^{\alpha}_{\beta}b_{i,j}\big|_{\beta=1}$ are, respectively, denoted as λ_1 and λ_2, which are determined by the data set and experiments.

Thirdly, the information degree (ID) vector of variable p and q with regard to the dependent variable α can be defined as

$${}^{\alpha}_{\beta}d_{p,q}\big|_{\beta=1} = {}^{\alpha}_{\beta}c_p\big|_{\beta=1} \odot {}^{\alpha}_{\beta}c_q\big|_{\beta=1} \tag{21}$$

where ${}^{\alpha}_{\beta}c_p\big|_{\beta=1}$ and ${}^{\alpha}_{\beta}c_q\big|_{\beta=1}$ is the pth and qth column vector of the matrix we calculate in Equation (20); $\odot$ represents Hadamard product, a binary operation that produces a $n \times 1$ matrix where each element is the product of elements in ${}^{\alpha}_{\beta}c_p\big|_{\beta=1}$ and ${}^{\alpha}_{\beta}c_q\big|_{\beta=1}$.

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

Table 2. Literature review of individual signature genes

Gene symbol	Full name	Reference	Description
PTBP1	Polypyrimidine tract binding protein 1	[32]	The expression of PTBP1 is related to platinum-resistance in OCa. PTBP1 plays a role in pre-mRNA splicing and in regulating alternative splicing events.
ANXA4	Annexin A4	[33]	ANXA4 is overexpressed in ovarian clear cell carcinoma and induces CR to platinum-based drugs.
OGN	Osteoglycin	[34]	OGN and PSAT1 are the major genes associated with cisplatin resistance in OCa.
GNG12	G protein subunit gamma 12	[35]	GNG12 is selected as chemotherapy-resistant gene of OCa.
WDR6	WD repeat domain 6	[36]	WDR6 is an upregulated signature gene in OCa data set with reduced sensitivity to platinum.
HLA-A	Major histocompatibility complex, class I, A	[37]	HLA-A is recognized as a down-regulated gene that is predominantly linked to the immune response in chemotherapy treatment and OCa.
IDO1	Indoleamine 2,3-dioxygenase 1	[38, 39]	IDO1 is positively associated with CR in paclitaxel-based therapy in OCa. IDO1 is also involved in the immune response.

The GIC ${}^{\alpha}_{\beta}\mathrm{GIC}_{p,q}\big|_{\beta=1}$ is defined as the percentage of non-zero elements in ID vector ${}^{\alpha}_{\beta}d_{p,q}\big|_{\beta=1}$:

$$ {}^{\alpha}_{\beta}\mathrm{GIC}_{p,q}\big|_{\beta=1} = \begin{cases} \frac{\mathrm{num}(1)}{n_\alpha} & \mathrm{num}(1) \geq \mathrm{num}(-1) \\ -\frac{\mathrm{num}(-1)}{n_\alpha} & \mathrm{num}(1) < \mathrm{num}(-1) \end{cases} \tag{22} $$

where num(1) and num(−1) is the number of element that equals to 1 and -1 in vector ${}^{\alpha}_{\beta}d_{p,q}\big|_{\beta=1}$.

Functional enrichment approaches

We performed functional analyses of our identified genes by using gene ontology (GO) and cell signaling pathways to evaluate the biological relevance and functional enrichment. All enrichment analyses were performed using the Enrichr https://amp.pharm.mssm.edu/Enrichr/ software tools [40, 41]. For cell signaling pathway analyses we used KEGG, WikiPathways, BioCarta and Reactome databases. We employed the Gene Ontology Biological Process database for gene ontological analysis [40]. For this work an adjusted *P*-value $\leq$0.05 was considered as statistically significant for functional analyses [42].

Evaluation

To evaluate the performance of signature gene pairs, a Support Vector Machine (SVM) classifier with linear kernel was selected to test the performance [43]. For training data $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$, the decision function of SVM is

$$ f(x) = sign(w^* \cdot x + b^*) \tag{23} $$

where $sign(\cdot)$ represents sign function and w^* and b^* is the optimal solution of this convex quadratic programming:

$$ \min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \qquad \text{s.t.} \quad \begin{cases} y_i(w \cdot x_i + b) \geq 1 - \xi_i, i = 1, 2, \ldots, n \\ \xi_i \geq 0, i = 1, 2, \ldots, n \end{cases} \tag{24} $$

By solving for the Lagrangian dual of the above problem, the simplified form is obtained as

$$ \min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{n}\alpha_i \qquad \text{s.t.} \quad \begin{cases} \sum_{i=1}^{n}\alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, n \end{cases} \tag{25} $$

When the optimal solution $\alpha^* = (\alpha_1{}^*, \alpha_2{}^*, \ldots \alpha_n{}^*)^{\mathrm{T}}$ is obtained, the key parameters w^* and b^* can be solved as

$$ \begin{cases} w^* = \sum_{i=1}^{n}\alpha_i{}^* y_i x_i \\ b^* = y_j - \sum_{i=1}^{n} y_i \alpha_i{}^* (x_i \cdot x_j), \ 0 < \alpha_i{}^* < C \end{cases} \tag{26} $$

Results

We applied our machine learning approach to identify GI related to OCa-CR. For this reason, we have preprocessed the TCGA data set. In order to reduce the complexity of the GI network, individual signature genes are identified based on the GBDT algorithm. In our experiments, we set experiment N = 20, and 35 individual signature genes are identified. Figures 2 and 3 show the frequency f_j of different genes in 20 experiments and the feature importance of individual signature genes in 20 experiments, respectively. Here, 35 individual signature genes are selected, as their frequency is higher than the threshold 10. The individual signature genes are validated using data processing indicators, and here Figure 4 presents the curves of data processing indicators (accuracy, precision and AUC) when the number of the individual signature genes increases gradually. It can be noticed that when the number of signature genes is more than 34, the data processing indicators reach the upper bound and hardly increase. In addition, 35 individual signature genes are enough to show a satisfactory prediction. These individual signature genes are finally selected to establish the GI network.

The 35 individual signature genes show satisfactory performance in the validation of data-processing indicators. Specially, the accuracy, precision and AUC reach 0.9410, 0.9309 and 0.9677, respectively. In addition to the above validation, we also seek

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

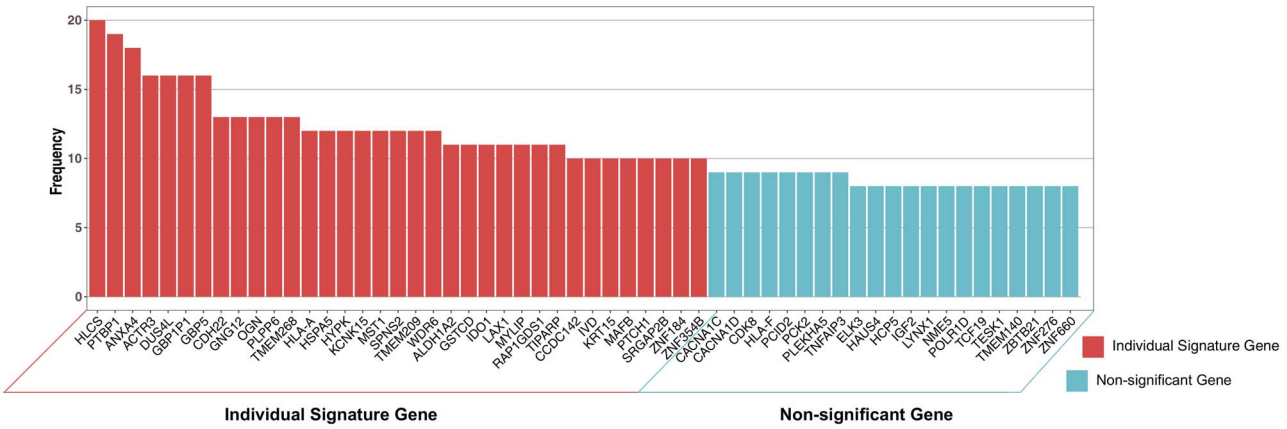

Figure 2. The identified individual signature genes and their frequency in 20 experiments.

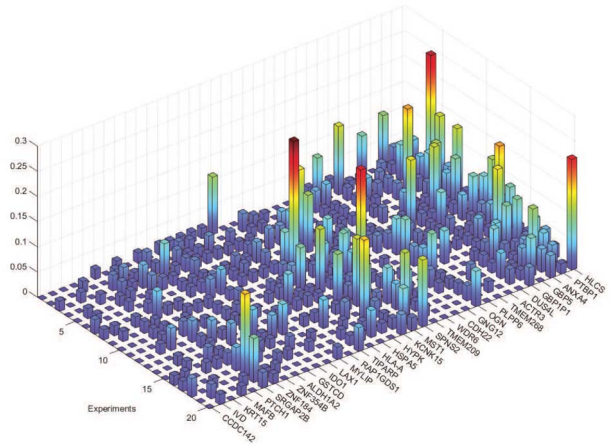

Figure 3. The feature importance of individual signature genes in 20 experiments.

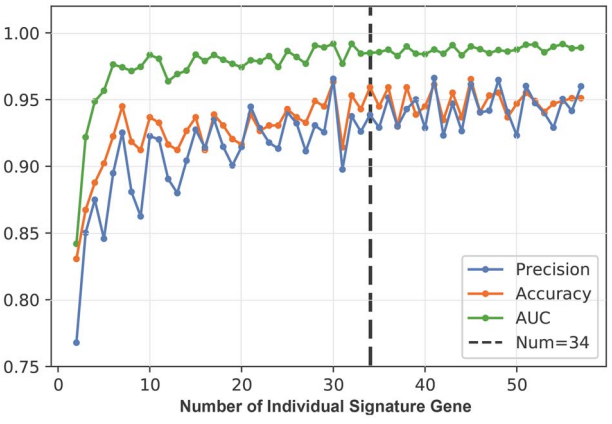

Figure 4. The accuracy, precision and AUC results when the number of individual signature gene increases gradually.

those individual signatures genes that have previously been reported in biological literature. We found that 7 of our 35 individual signature genes have previously been confirmed as being associated with OCa-CR. These 7 individual signature genes are PTBP1, ANXA4, OGN, GNG12, WDR6, HLA-A and IDO1, and their details are shown in Table 2. In addition to seven signature genes

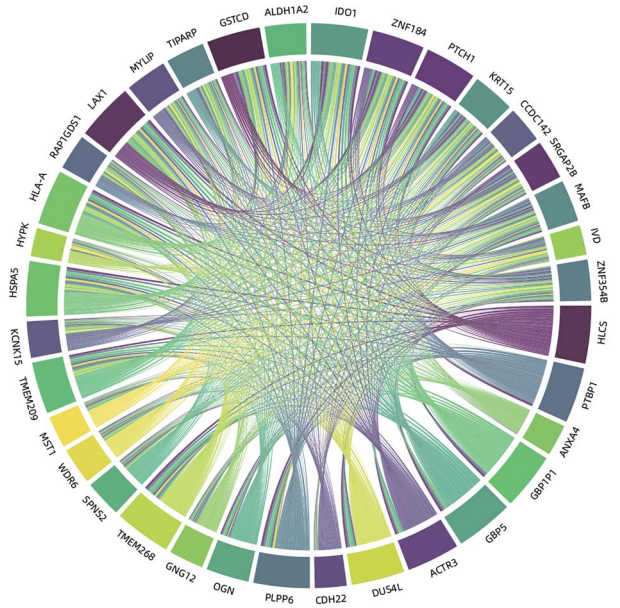

Figure 5. The GI network of selected 35 individual signature genes, where the width of connection represents GIC value and the size of the node represents frequency f_j in Equation (10).

that are supported by literature evidence, other signature genes are newly discovered, which may lead to further enlightenment on future research of OCa-CR.

As presented above, GIC measures the correlation among genes. Here, GIC can be calculated for totally 595 gene pairs based on the selected 35 individual signature genes. In Figure 5, a GI network of individual signature genes is shown, where the width of connection denotes the GIC value and the size of the node represents frequency f_j in Equation (10).

It can be seen that a small number of gene pairs show significant GIC values. Here, we let α represents CR, and we also set $\lambda_1 = 0.2$ and $\lambda_2 = -0.2$. We set the threshold of GIC as 0.75 to select the signature gene pairs, and totally 24 gene pairs are identified. These signature gene pairs form a simplified GI network, which is presented in Figure 6. Each edge represents a signature gene pair, and GIC is the width of the edge. This network is easy to understand since the number of individual signature genes is rather small and only the important GI is shown.

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

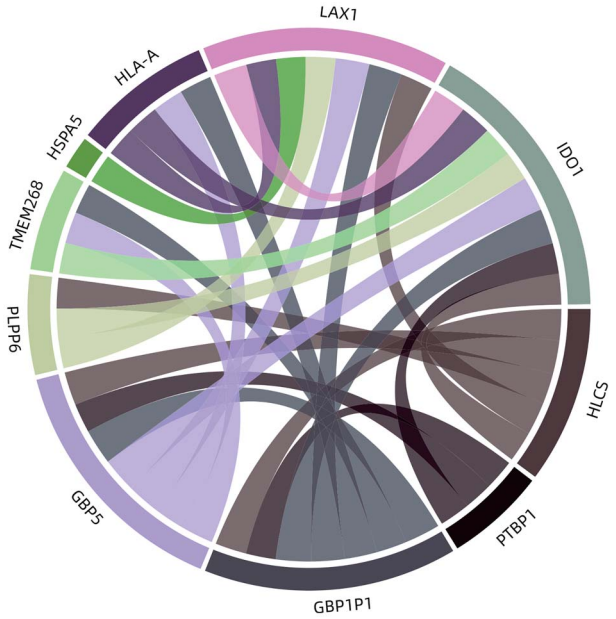

Figure 6. The GI network of 24 signature gene pairs, where the width of connection represents GIC value and the size of node represents frequency f_j in Equation (10). This network is easy to understand since the number of individual signature genes is rather small and only the important GI is shown.

Table 3. The number of connections of each signature gene

Signature gene	Number of connections
GBP1P1	7
GBP5	7
IDO1	8
HLA-A	4
HLCS	5
HSPA5	1
LAX1	7
PLPP6	3
PTBP1	3
TMEM268	3

From Figure 6, we found that each signature gene has different number of connections with other signature genes. The number of connections of each signature gene is quantified in Table 3. These connections may help the researchers to delineate the pathways, protein complexes and underlying biological processes of OCa-CR. For example, the number of connections of signature gene IDO1 is 8, which implies that IDO1 and its gene interactions are important factors in underlying biological mechanism.

The identified signature gene pairs are validated by information value, where the latter is an index that measures a feature's influence on the target [44]. The information values of signature gene pairs are presented in Figure 7, where different levels of information value are presented in different colors. A feature is considered to be predictive when its information value is larger than 0.3. We can notice that these signature gene pairs show satisfactory performance on information value validation.

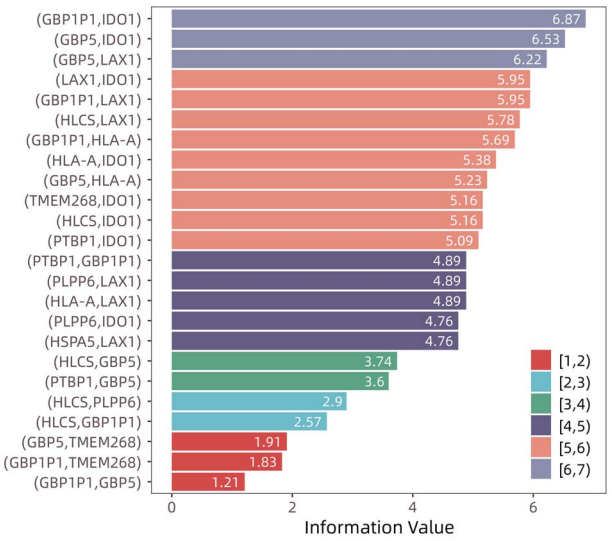

Figure 7. The information value of selected signature gene pairs, where the high information value represents signature gene pair's satisfying predicting ability. Here different levels of information value are presented in different colors.

Functional enrichment

We have performed the functional analyses of our identified genes. In this regard we have performed signaling pathway and GO analysis as shown in Figure 8 and 9. We found that our identified biomarkers are strongly associated with the cancer immune-related pathways including proteins with altered expression in cancer immune redescape, transcription factors in beta–cell neogenesis and MHC1 causes antigen presentation failure in cancer immune escape signaling pathways that are strongly associated with cancer progression. Similarly, we found several functional (biological) pathways that are strongly associated with the OCa progression from the GO and functional pathway analysis.

Discussions and conclusion

In this study, we investigated the GI related to OCa-CR. We employed the GBDT approach to demonstrate a method for better reducing the dimension in the gene network. Thus, 35 individual signature genes are selected among 14252 genes. These individual signature genes show satisfactory performance in the validation of data-processing indicators, where the accuracy, precision and AUC are 0.9410, 0.9309 and 0.9677, respectively. Here we also found 7 genes, PTBP1, ANXA4, OGN, GNG12, WDR6, HLA-A and IDO1 out of our 35 identified individual signature genes are confirmed in previously published biological literature. Since GIC method measures the correlation to denote association between two quantitative variables and explain two independent variables with joint effect on the dependent variable, it seems natural to choose GIC over other approaches of identification of the signature gene pairs on OCa-CR. In our study, totally 24 signature gene pairs with high GIC weights are identified, and their information values are calculated as well. It is interesting to notice that 24 signature gene pairs include 10 individual genes only. The identified signature gene pairs not only build an efficient GI network of OCa-CR but provide an alternative way to predict the OCa-CR as well.

We have also applied bubble heat chart to visualize the average target value across interaction between signature gene pairs.

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

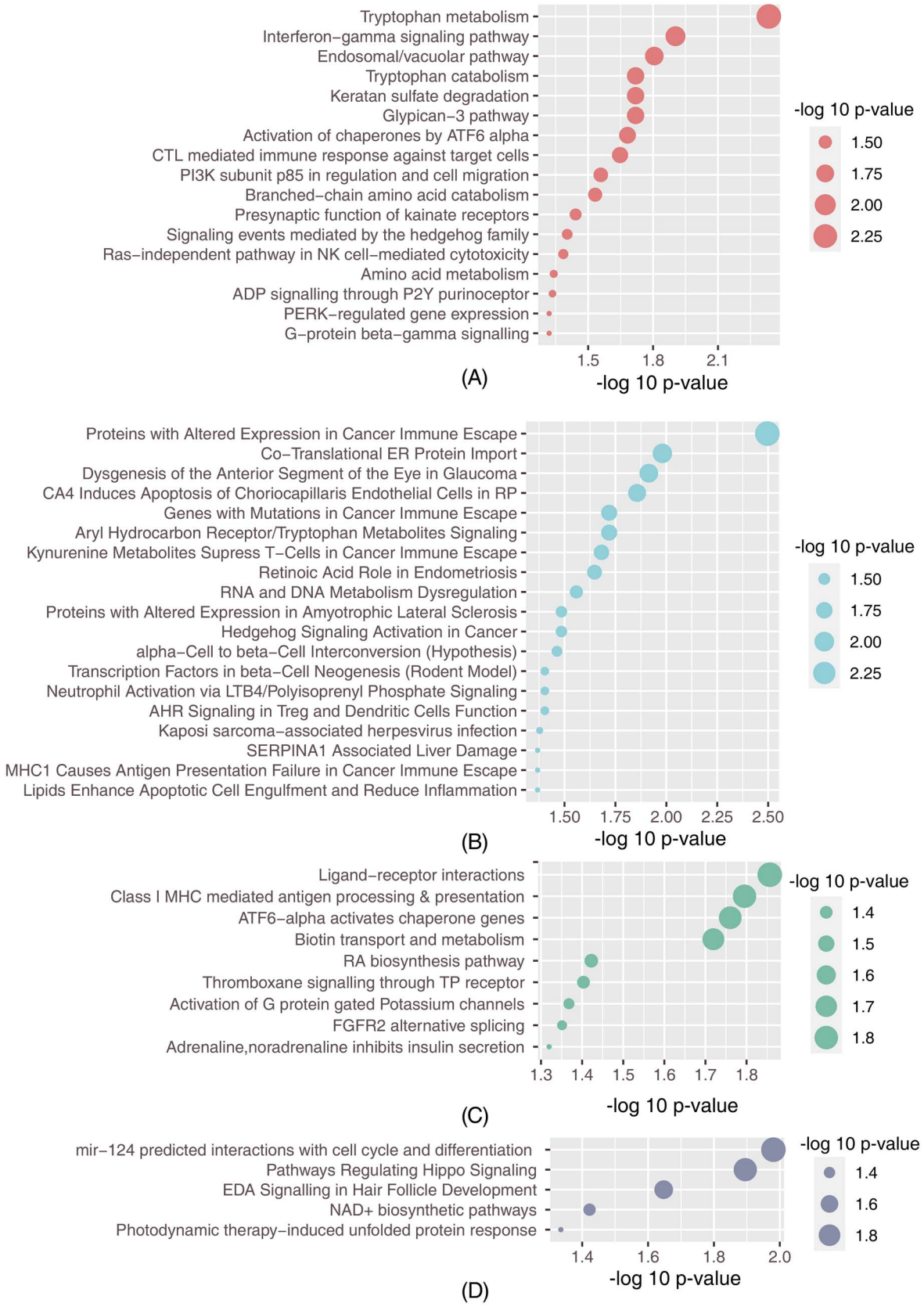

Figure 8. Significant signaling pathways associated with the identified genes. A.Top significant pathways using the Bioplanet pathway database B.Top significant pathways using the KEEG pathway database C. Top significant pathways using the REACTOM pathway database D. Top significant pathways using the WikiPathways pathway database.

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

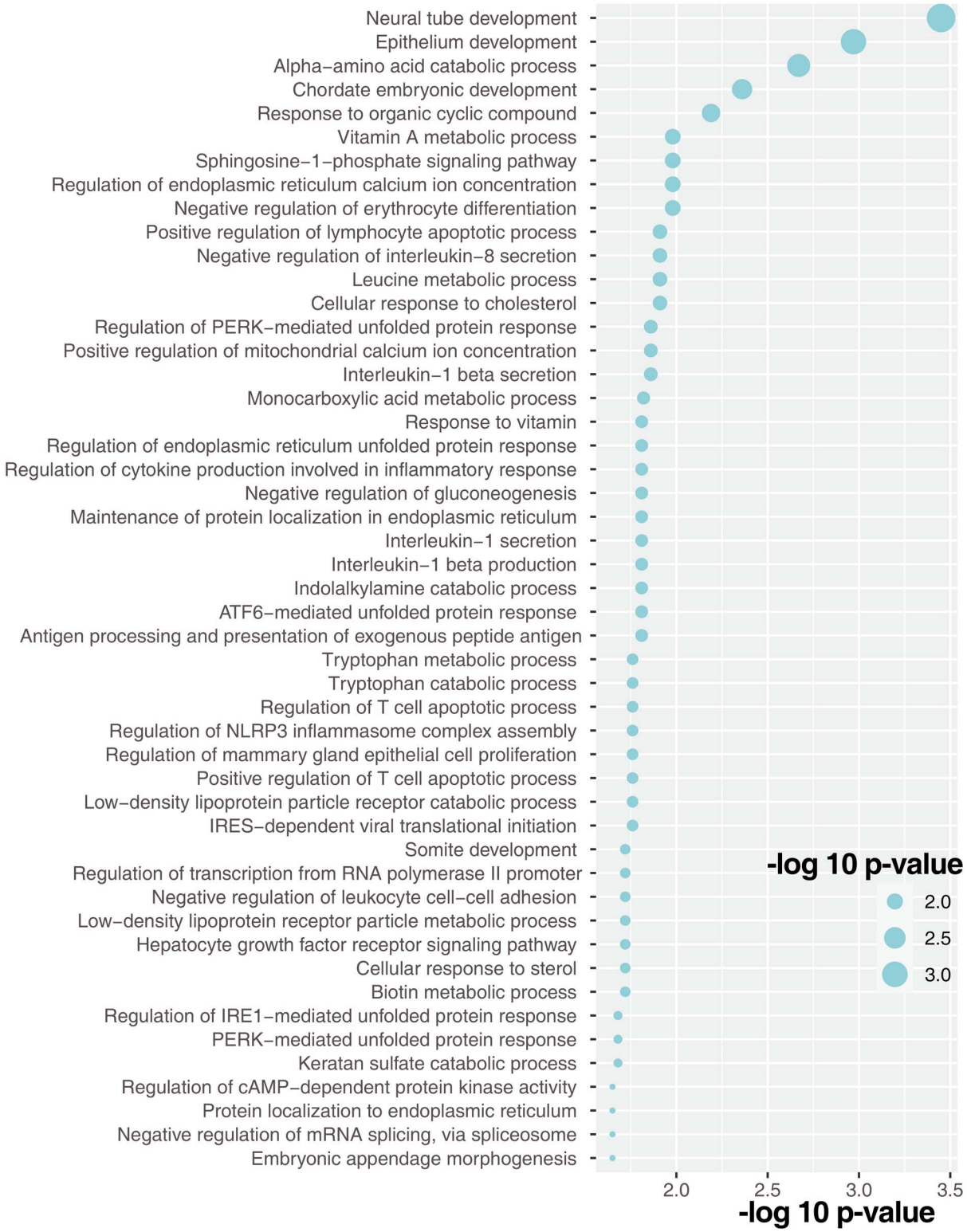

Figure 9. Significant GO (biological process) pathways associated with the identified genes.

Figure 10 presents a positive relationship in gene pair (num1 and GBP5). If we consider gene pair as a composite feature, then GIC helps to find out these composite features that differentiate OCa-CR and OCa-CS patients. Figure 10 is a visualization of how signature gene pairs affect OCa-CR [45]. The most important insight comes from the color of the bubble, darker color represents a higher probability of CR. The size of the bubble implies the number of samples in that class. For simplicity we use IDO1 and GBP5 to substitute ${}^{\alpha}_{\beta}c_{i,IDO1}\big|_{\beta=1}$, ${}^{\alpha}_{\beta}c_{i,GBP5}\big|_{\beta=1}$ in Equation (20). From Figure 10, we can notice that patient i has the highest

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

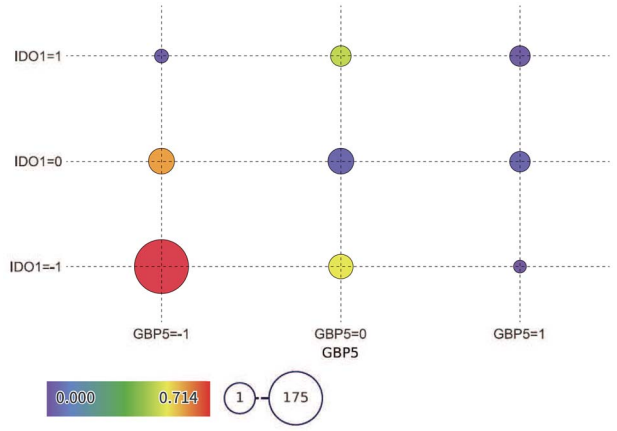

Figure 10. Bubble heat map uses a warm-to-cool color spectrum, the warm color of the bubble represents higher probability of OCa-CR. This figure visualizes the OCa-CR distribution in gene pairs that possess positive relationship. (IDO1 and GBP5) is taken as an example here. Patient *i* has the highest probability to show CR when IDO1 = −1 and GBP5 = −1.

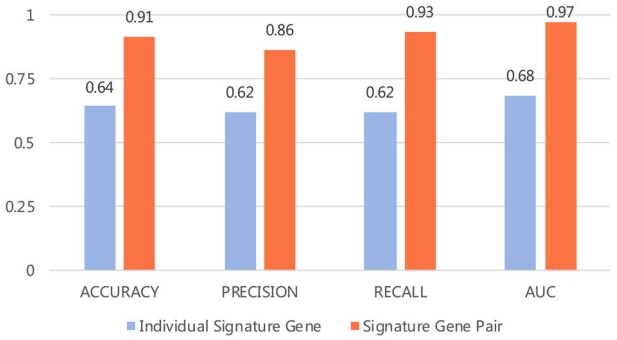

Figure 11. The accuracy, precision, recall and AUC results of 24 signature gene pairs, compared to the results of 10 individual signature gene.

probability to show CR when IDO1 = −1 and GBP5 = −1. If patient *i* only satisfies one condition, rather than both two conditions, the probability of CR will decrease.

It can be noticed from the GI network in Figure 6 that 24 signature gene pairs comprise of 10 signature genes only. SVM classifier with a linear kernel is applied to compare the performance of signature gene pairs and individual signature genes. As shown in Figure 11, the accuracy, precision, recall and AUC of signature gene pairs reach 0.9082, 0.8636, 0.9268 and 0.9658, respectively. On the other hand, the accuracy, precision, recall and AUC of 10 individual signature genes are 0.6367, 0.6167, 0.6244 and 0.6823, respectively. Apparently, the predicting performance of signature gene pairs is much better than individual signature genes. In summary, we regard our proposed methods as a useful tool to investigate the GI related to OCa-CR. Our method could be useful to develop potential therapies, and patients with OCa may expect a better genetic diagnosis in future. Moreover, using this identified biomarker, it is possible to predict OCa patients at the early stage and these prognostic markers could be useful at the genetic clinics for the diagnosis.

Code Availability

All of the methods are implemented in Python. The original data and source code is available at GitHub page: https://github.com/Nikki0526/gene-pair-research.

Author contributions

Y. Lei conceptualized the study; K. Chen and Y. Lei developed the composite analysis method; K. Chen, H. Xu, Y. Lei and M. Moni performance the coding and data analysis; K. Chen, H. Xu, Y. Lei and M. Moni prepared the figures; Y. Li and H. Guo organized the downloaded data; and K. Chen, H. Xu, Y. Lei, M. Moni and P. Lio wrote this paper.

Key Points

- This work investigates the genetic interaction related to ovarian cancer (OCa) chemoresistance (CR).
- The individual signature genes are selected using a machine learning approach.
- The gene interaction coefficient is proposed to identify the signature gene pairs on OCa-CR and provided a signature pair-based prediction method of OCa-CR.
- This work may lead to the discovery of possible gene interaction related to OCa-CR.

Funding

This work is partially supported by Ningxia Key Research and Development Program (Grant No. 2019BFG02002).

Acknowledgements

The authors thank the editor and anonymous reviewers for their valuable suggestions. The authors wish to acknowledge the support from School of Electronics Engineering and Computer Science, Peking University, Beijing, China, the Department of Biomedical Engineering, Duke University, Durham, USA, the School of Public health and Community Medicine, University of New South Wales, Sydney, Australia, the Computer Laboratory, University of Cambridge, Cambridge, UK, and the Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China.

References

1. Lønning PE, Knappskog S. Mapping genetic alterations causing chemoresistance in cancer: identifying the roads by tracking the drivers. *Oncogene* 2013;**32**(46):5315–30.
2. Xu H, Moni MA, Liò P. Network regularised cox regression and multiplex network models to predict disease comorbidities and survival of cancer. *Comput Biol Chem* 2015;**59**(1):15–31.
3. Dorman SN, Baranova K, Knoll JH, *et al.* Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Mol Oncol* 2016;**10**(1):85–100.
4. Fersini E, Messina E, Leporati A. Discovering gene-drug relationships for the pharmacology of cancer. *Adv Comput Intell* 2012;**14**(1):117–26.
5. Wani S, Drahos J, Cook MB, *et al.* Comparison of endoscopic therapies and surgical resection in patients with early esophageal cancer: a population-based study. *Gastrointest Endosc* 2014;**79**(2):224–32.
6. Tan AC, Naiman D, Xu L, *et al.* Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 2005;**21**(1):3896–904.

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024

7. Goddard RU, Eccles D, Fliege J, *et al*. Machine learning approaches for the discovery of gene–gene interactions in disease data. *Brief Bioinform* 2012;**14**(2):251–60.
8. Cordell H. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;**10**(1):392–404.
9. Park M, Lee J, Park T, *et al*. Gene-gene interaction analysis for the survival phenotype based on the Kaplan-Meier median estimate. *Biomed Res Int* 2020;**20**(1):1–10.
10. Ideker T, Krogan NJ. Differential network biology. *Mol Syst Biol* 2012;**8**(1):10–22.
11. Yulan D, Shangyi L, Chunyu D, *et al*. Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. *Brief Bioinform* 2017;**13**(1):1.
12. Moni MA, Liò P. Network-based analysis of comorbidities risk during an infection: SARS and HIV case studies. *BMC bioinformatics* 2014;**15**(1):1–23.
13. Wildenhain J, Spitzer M, Dolma S, *et al*. Prediction of synergism from chemical-genetic interactions by machine learning. *Cell Systems* 2015;**1**(6):383–95.
14. Chirigati F, Doraiswamy H, Damoulas T, *et al*. Data polygamy: The many-many relationships among urban spatio-temporal data sets. In: *Proceedings of the 2016 International Conference on Management of Data, page*, 2016, 1011–25.
15. Hao C, Yong H, Jiadong J, *et al*. A machine learning method for identifying critical interactions between gene pairs in alzheimer's disease prediction. *Front Neurol* 2019;**10**: 1162.
16. Madhukar NS, Elemento O, Pandey G. Prediction of genetic interactions using machine learning and network properties. *Front Bioeng Biotechnol* 2015;**3**(1):172.
17. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 2001;**29**(5):1189–232.
18. Li Q, Zhao K, Carlos D, , *et al*.. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genet Med* 2019;**21** (9): 2126 –34.
19. Moni MA, Rana HK, Islam MB, *et al*. A computational approach to identify blood cell-expressed parkinson's disease biomarkers that are coordinately expressed in brain tissue. *Comput Biol Med* 2019;**113**(1):103–385.
20. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;**474**(7353):609–15.
21. Davide R, Katja S, Gavin S, *et al*. Gc-content normalization for rna-seq data. *BMC Bioinformatics* 2011;**12**(1):480.
22. Kuhn M. Building predictive models in r using the caret package. *J Stat Softw* 2008;**28**(1):1–26.
23. Liu C. Classifying dna methylation imbalance data in cancer risk prediction using smote and tomek link methods. *International Conference of Pioneering Computer Scientists, Engineers and Educators* 2019.
24. Chawla NV, Bowyer KW, Hall LO, *et al*. Smote: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**(1):321–57.
25. Maniruzzaman M, Rahman MJ, Ahammed B, *et al*. Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. *Comput Methods Programs Biomed* 2019;**176**(1):173–93.
26. Zengqiang F, Xiujuan L. GBDTCDA: predicting circRNA-disease associations based on gradient boosting decision tree with multiple biological data fusion. *Int J Biol Sci* 2019;**15**(1):2911–24.
27. Li W, Zhang W, Zhang J. A novel model integration network inference algorithm with clustering and hub genes finding. *Molecular informatics* 2020;**39**(5):190.
28. Taninaga J, Nishiyama Y, Fujibayashi K, *et al*. Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: a case-control study. *Sci Rep* 2019;**9**(1):12384.
29. Stephan J, Stegle O, Beyer A. A random forest approach to capture genetic effects in the presence of population structure. *Nat Commun* 2015;**6**(1):7432.
30. Sweeney Y, Clopath C. Population coupling predicts the plasticity of stimulus responses in cortical circuits. *Elife* 2020;**9**(1):99.
31. Mignan MBA. One neuron versus deep learning in aftershock prediction. *Nature* 2019;**574**:E1–3.
32. Severi L, Losi L, Fonda S, *et al*. Proteomic and bioinformatic studies for the characterization of response to pemetrexed in platinum drug resistant ovarian cancer. *Front Pharmacol* 2018;**9**(1):454.
33. Kim A, Enomoto T, Serada S, *et al*. Enhanced expression of annexin a4 in clear cell carcinoma of the ovary and its association with chemoresistance to carboplatin. *Int J Cancer* 2009;**7**(10):452–2.
34. Dai J, Wei R, Zhang P, *et al*. Overexpression of microrna-195-5p reduces cisplatin resistance and angiogenesis in ovarian cancer by inhibiting the psat1-dependent gsk3 β/β-catenin signaling pathway. *J Transl Med* 2019;**17**(1):190.
35. Sun J, Bao S, Xu D, *et al*. Large-scale integrated analysis of ovarian cancer tumors and cell lines identifies an individualized gene expression signature for predicting response to platinum-based chemotherapy. *Cell death and disease* 2019;**10**(9):661.
36. Koussounadis A, Langdon S, Harrison D, *et al*. Chemotherapy-induced dynamic gene expression changes in vivo are prognostic in ovarian cancer. *Br J Cancer* 2014;**110**(12):2975–84.
37. L'Espérance S, Bachvarova M, Tetu B, *et al*. Global gene expression analysis of early response to chemotherapy treatment in ovarian cancer spheroids. *BMC Genomics* 2008;**9**(1):99.
38. Okamoto A, Nikaido T, Ochiai K, *et al*. Indoleamine 2,3-dioxygenase serves as a marker of poor prognosis in gene expression profiles of serous ovarian cancer cells. *Clin Cancer Res* 2005;**11**(16):6030–9.
39. Zhai ALL, Ladomersky E. Ido1 in cancer: a gemini of immune checkpoints. *Cellular and Molecular Immunology* 2018;**10**(15):447–57.
40. Kuleshov MV, Jones MR, Rouillard AD, *et al*. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**(W1):W90–7.
41. Moni MA, Liò P. comoR: a software for disease comorbidity risk assessment. *Journal of clinical bioinformatics* 2014;**4**(1):1–11.
42. Moni MA, Liò P. How to build personalized multi-omics comorbidity profiles. *Frontiers in cell and developmental biology* 2015;**3**(1):28.
43. Lopez R, Wang R, Seelig G. A molecular multi-gene classifier for disease diagnostics. *Nat Chem* 2018;**10**(1):746–54.
44. Howard RA. Information value theory. *IEEE Transactions on Systems Science and Cybernetics* 1966;**2**(1):22–6.
45. Goldstein A, Kapelner A, Bleich J, *et al*. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat* 2015;**24**(1):44–65.

Downloaded from https://academic.oup.com/bib/article/22/6/bbab100/6272796 by guest on 18 December 2024