Received: 26 September 2023 | Accepted: 21 November 2023

DOI: 10.1002/cai2.110

ORIGINAL ARTICLE

CANCER INNOVATION

Novel progressive deep learning algorithm for uncovering multiple single nucleotide polymorphism interactions to predict paclitaxel clearance in patients with nonsmall cell lung cancer

Wei Chen[1,2] | **Haiyan Zhou**[2] | **Mingyu Zhang**[2] | **Yafei Shi**[2] | **Taifeng Li**[2] | **Di Qian**[2] | **Jun Yang**[2] | **Feng Yu**[1] | **Guohui Li**[2]

[1]School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, China

[2]Pharmacy Department, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Correspondence

Jun Yang and Guohui Li, Pharmacy Department, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, No. 17 Panjiayuan St South, Beijing 100021, China.
Email: yangjun@cicams.ac.cn and lgh0603@cicams.ac.cn

Feng Yu, School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, No. 639 Longmian Dadao, Nanjing 211198, China.
Email: yufeng@cpu.edu.cn

Funding information

Beijing Hope Run Special Fund of the Cancer Foundation of China, Grant/Award Number: LC2020L03; CAMS Innovation Fund for Medical Sciences, Grant/Award Number: 2021-I2M-1-014

Abstract

Background: The rate at which the anticancer drug paclitaxel is cleared from the body markedly impacts its dosage and chemotherapy effectiveness. Importantly, paclitaxel clearance varies among individuals, primarily because of genetic polymorphisms. This metabolic variability arises from a nonlinear process that is influenced by multiple single nucleotide polymorphisms (SNPs). Conventional bioinformatics methods struggle to accurately analyze this complex process and, currently, there is no established efficient algorithm for investigating SNP interactions.

Methods: We developed a novel machine-learning approach called GEP-CSIs data mining algorithm. This algorithm, an advanced version of GEP, uses linear algebra computations to handle discrete variables. The GEP-CSI algorithm calculates a fitness function score based on paclitaxel clearance data and genetic polymorphisms in patients with nonsmall cell lung cancer. The data were divided into a primary set and a validation set for the analysis.

Results: We identified and validated 1184 three-SNP combinations that had the highest fitness function values. Notably, *SERPINA1*, *ATF3* and *EGF* were found to indirectly influence paclitaxel clearance by coordinating the activity of genes previously reported to be significant in paclitaxel clearance. Particularly intriguing was the discovery of a combination of three SNPs in genes *FLT1*, *EGF* and *MUC16*. These SNPs-related proteins were confirmed to interact with each other in the protein–protein interaction network, which formed the basis for further exploration of their functional roles and mechanisms.

Abbreviations: GEP, gene expression programming; PPI, protein-protein interaction; SNP, single nucleotide polymorphism.

Conclusion: We successfully developed an effective deep-learning algorithm tailored for the nuanced mining of SNP interactions, leveraging data on paclitaxel clearance and individual genetic polymorphisms.

KEYWORDS
deep learning algorithm, paclitaxel, single nucleotide polymorphisms

1 | INTRODUCTION

Single nucleotide polymorphisms (SNPs) are variations at specific positions in the genome where a single nucleotide differs. They can serve as valuable markers for pinpointing genetic variations correlated with clinical phenotypes. Identifying SNPs that are associated with clinical phenotypes is a pivotal task in genome-wide association studies. However, analyzing complex clinical phenotypes is a significant challenge. The complexity is exacerbated by potential interactions between phenotypes and SNPs [1, 2]. One strategy has involved the integration of interaction terms in statistical models [3, 4]. Such integration can elucidate the possible influences of SNPs on a phenotype, thereby refining the precision of the analysis. However, alternative techniques such as machine-learning algorithms may offer a robust mechanism to discern intricate interplays between variables [5, 6].

Paclitaxel is an important and effective anticancer drug [7] that is used to treat solid tumors, including breast cancer [8], lung cancer [9, 10] and ovarian cancer [11]. The efficacy of paclitaxel varies among individuals because of genetic disparities that arise chiefly from SNPs. Several genes and their translated proteins, which regulate cell behavior, have been shown to influence the pharmacokinetics, pharmacodynamics [12] and toxicity [13] of paclitaxel. Nevertheless, the mechanism underlying the therapeutic effect of paclitaxel [14] is intricate and not fully understood. Clinical phenotypes are frequently shaped by the combined effects of multiple genes [15], and therefore alterations in a single gene often fail to capture the complexity of clinical phenotypes.

SNPs that are located in genes that encode drug transporters, metabolizing enzymes, or drug targets can have an impact on the pharmacokinetics of paclitaxel. Consequently, SNPs have the potential to cause divergent responses in drug efficacy and toxicity. Clearance, a pivotal pharmacokinetic metric, indicates the rate of drug elimination from the body. Notably, paclitaxel has nonlinear clearance patterns at high doses [16]. Understanding clearance is essential for determining optimal dosing regimens and curtailing the likelihood of adverse reactions. Identifying SNP combinations that influence paclitaxel clearance may help clinicians identify individuals who might have altered drug metabolism or responses, thereby enabling the implementation of a personalized treatment approach. Such an approach can also deepen the understanding of underlying biological pathways and potentially contribute to the development of novel therapeutic strategies, including targeted drug delivery and the use of adjuvant medications to augment paclitaxel efficacy. Nonetheless, critical questions remain, including which specific SNP combinations impact paclitaxel clearance, and what methodologies can be used to discover these combinations. Current methods do not conclusively analyze the interplay between drug clearance and SNP interactions [2, 17]. We have developed a novel machine-learning approach called the gene expression programming-complex snp interactions (GEP-CSI) data mining algorithm. This algorithm, an evolution of GEP [18], leverages linear algebra calculations to handle discrete variables and compute a fitness function score. GEP enables the exploration of an extensive solution space, pinpointing the optimal solution via a natural selection mechanism within the algorithm.

The GEP-CSI algorithm facilitates the development of bespoke functions and operators tailored to optimize specific SNP combinations, with a particular focus on high clearance. By applying this algorithm to complex phenotypes, we identified potential SNP interactions associated with high paclitaxel clearance. These interactions warrant further examination in future studies.

2 | METHODS

2.1 | Study design

In this study, we aimed to develop a deep machine-learning algorithm for investigating multiple SNP interactions. The study was conducted at the National Cancer Center/Cancer Hospital affiliated with the Chinese Academy of Medical Sciences and Peking

Union Medical College in Beijing, China. Ethical principles outlined in the Declaration of Helsinki were strictly adhered to, and the study was approved by the Institutional Review Board of the Cancer Hospital, Chinese Academy of Medical Sciences (approval number: 15-123/1050). The clinical trial was registered with the Chinese Clinical Trial Registry (www.chictr.org.cn) under registry number ChiCTR2000040300.

2.2 | Patients

A single-center clinical trial was conducted on 30 patients who were treated with paclitaxel at the Cancer Hospital of the Chinese Academy of Medical Sciences between June 2015 and August 2018. This trial focused on the pharmacokinetics, pharmacodynamics, and pharmacogenetics of paclitaxel. Each patient was diagnosed with squamous cell nonsmall cell lung cancer, which was confirmed by cytology or histology. Eligibility criteria were: age 18–70 years; anticipated survival period >3 months; no radiation or other chemotherapy treatments in the prior 3 weeks; adequate hematopoietic function evidenced by absolute neutrophil counts $>1.5 \times 10^9\,L^{-1}$ and platelet counts $>100 \times 10^9\,L^{-1}$; satisfactory liver and kidney function as indicated by alanine transaminase, aspartate aminotransferase, total bilirubin, and creatinine levels all <1.5 times the upper normal limit; and no consumption of inducers or inhibitors of liver metabolic enzymes in the previous 4 weeks. Patients with active infections, severe medical conditions, or other tumor types were excluded from the study.

2.3 | Purification of DNA

DNA was purified from plasma samples as follows. Briefly, plasma (1 mL) was mixed with 800 μL Buffer ACL, and 100 μL protease K, and incubated at 60°C for 30 min. Then, the mixture was combined with 1.8 mL Buffer ACB and left on ice for 5 min. Subsequently, the buffer was drawn through a QIAamp MinElute column inserted into the VacConnector on the vacuum manifold. Buffer ACW1 (600 μL), Buffer ACW2 (750 μL), and absolute ethanol were added sequentially to the column, with each buffer drawn through using a vacuum pump. The column was centrifuged at $18{,}000 \times g$ for 3 min, followed by incubation at 56°C for 10 min to ensure complete membrane drying. Buffer AVE (20–150 μL) was added to the column, which was then incubated at room temperature for 3 min. After centrifugation at $18{,}000 \times g$ for 1 min, the DNA was transferred to new tubes and stored at −80°C. The buffers used in the purification were taken from the QIAamp Circulating Nucleic Acid Kit (Qiagen).

2.4 | Whole exome sequencing

High-quality genomic DNA samples were fragmented randomly using an ultrasonicator (Covaris, LLC. Woburn). We selected 150–250 bp fragments and performed end-repair of the DNA fragments, added an "A" base to the 3′ end, and attached library adapters to both ends. This was followed by ligation-mediated PCR amplification of the ligated library to generate a hybridization library. A substantial portion of the hybridization library was captured and enriched using an exon array. Unenriched fragments were discarded before the amplification process. The amplified products were separated into single-strand and circularized fragments. The circularized library underwent rolling circle amplification resulting in DNA nano balls (DNBs). The DNA nano balls that met the quality control criteria were sequenced on a BGISEQ-500 platform (BGI), ensuring that the data volume met the prescribed requirements. The resultant paired-end reads were stored in FASTQ format as raw data.

2.5 | Bioinformatics analysis

The raw data were filtered to remove adapter sequences, low-quality bases, and unsequenced bases to obtain clean reads. The clean data were aligned to the human reference genome using the Burrows-Wheeler Aligner software. Duplicate reads were removed with Picard tools, and local realignment and base quality recalibration were performed using the Genome Analysis Toolkit. The aligned sequences were analyzed using evaluation metrics, including sequencing depth, coverage, and alignment rate for each sample. A rigorous data quality control system was enforced consistently throughout the entire analysis pipeline to ensure that high-quality sequencing data were obtained.

The HaplotypeCaller module from the Genome Analysis Toolkit v3.7 was used in the pipeline to detect genomic variations, specifically SNPs and insertion/deletions. The initial variants were filtered to obtain high-confidence variant data. Subsequently, SnpEff software (http://snpeff.sourceforge.net/SnpEff_manual.html) was used to annotate and predict the effects of these variants

for subsequent analysis. The bioinformatics workflow is shown in Figure 1.

2.6 | Exon data preprocessing

The results of the pharmacokinetic analysis of paclitaxel in the 30 patients were used to categorize the patients as high-clearance or low-clearance. For the subsequent analysis in R, all the SNP values and phenotypes were transformed into binary variables scaled to 0 and 1.

2.7 | Statistical analysis

Blood concentration data were from a previous study [19]. Paclitaxel clearance was computed using NONMEM v7.3. The model was assessed using the Stepwise Covariate Model module in Perl-speaks-NONMEM software (v3.2.12). Pearson's chi-square test was used to evaluate deviations from the null hypothesis, which posited that patients in both the high- and low-clearance groups had identical genotype count distributions. Statistical evaluation of the relationship between a specific SNP and clearance was conducted using R software. SNPs with $p < 0.01$ were retained.

2.8 | Development of a deep-learning method for SNP interactions

The GEP-CSI algorithm was devised using GEP and implemented in Python in a virtual environment on the Deepin operating system, which is based on the Linux kernel. Designed to discover multiple SNP interactions underlying the phenotypes, the algorithm operated progressively, scaling as the number of SNPs increased. This stepwise strategy was also reflected in

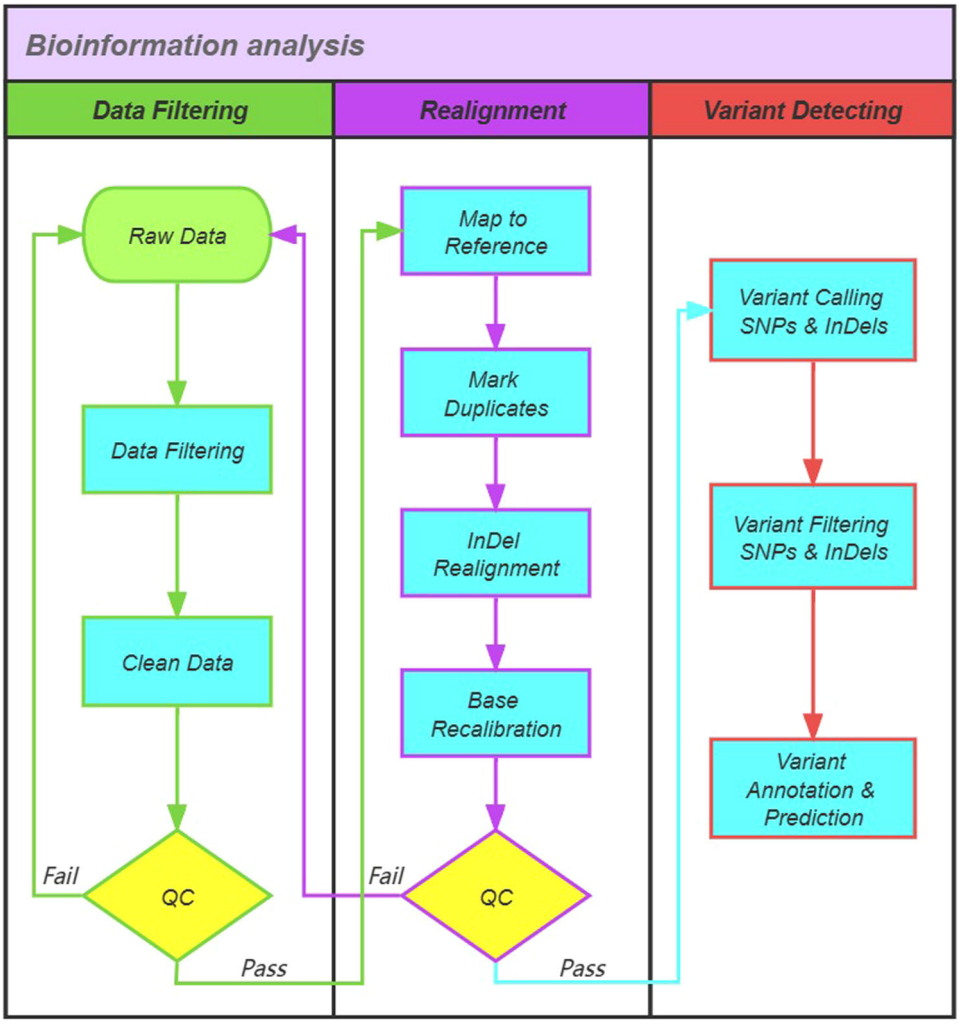

FIGURE 1 Bioinformatics workflow used in this study. InDel, insertion and deletion; QC, quality control; SNP, single nucleotide polymorphism.

variations in gene population initialization parameters. For instance, for an interaction involving two SNPs, the parameters were gene number 2, number of linking genes 1, gene head length 2, and linking gene head length 3, and for an interaction involving three SNPs, the parameters were gene number 3, number of linking genes 1, gene head length 4, and linking gene head length 5.

For each SNP combination, the GEP-CSI algorithm processed 100 generations, each housing a population of 20 chromosomes, to fine-tune the gene expression and attain the optimal solution. The function symbol set incorporated three logical operators, 'and', 'or', and 'not'. The count of variables in the terminator set was contingent on the number of SNPs in the given combination. The fitness function used matrix computation methods from linear algebra, where 'A' is the detection result of an SNP and 'B' is the metabolic outcome of paclitaxel clearance. Both 'A' and 'B' are one-dimensional matrices populated by 1 and 0. To toggle matrix elements between 0 and 1, α and β were set as -1 and 1, respectively.

TABLE 1 Characteristics of patients in the high- and low-clearance groups.

Characteristic	All, $N = 30$	High clearance ($N = 15$)	Low clearance ($N = 15$)	p-value
Age				0.65
Mean (SD)	58 (9)	58 (9)	57 (9)	
Median (IQR)	60 (54, 64)	60 (54, 64)	55 (54, 64)	
Range	37, 75	38, 75	37, 71	
Gender, *n* (%)				>0.99
Female	4 (13.3)	2 (13.3)	2 (13.3)	
Male	26 (86.7)	13 (86.7)	13 (86.7)	
Smoking history, *n* (%)				>0.99
No	5 (16.7)	2 (13.3)	3 (20.0)	
Yes	25 (83.3)	13 (86.7)	12 (80.0)	
Drinking history, *n* (%)				0.72
No	15 (50.0)	8 (53.3)	7 (46.7)	
Yes	15 (50.0)	7 (46.7)	8 (53.3)	
T stage, *n* (%)				0.64
T1	3 (10.0)	2 (13.3)	1 (6.7)	
T2	12 (40.0)	5 (33.3)	7 (46.7)	
T3	8 (26.7)	3 (20.0)	5 (33.3)	
T4	6 (20.0)	4 (26.7)	2 (13.3)	
X	1 (3.3)	1 (6.7)	0 (0.0)	
N stage, *n* (%)				0.43
N0	2 (6.7)	2 (13)	0 (0.0)	
N1	4 (13.3)	3 (20)	1 (6.7)	
N2	17 (56.7)	7 (47)	10 (66.6)	
N3	6 (20.0)	3 (20)	3 (20.0)	
X	1 (3.3)	0 (0)	1 (6.7)	
M stage, *n* (%)				0.14
M0	21 (70.0)	13 (87)	8 (53.4)	
M1	7 (23.3)	2 (13)	5 (33.3)	
X	2 (6.7)	0 (0)	2 (13.3)	

Note: Age: Wilcoxon rank sum test; Others: Fisher's exact test.

$$\text{fitness} = \frac{\{(A * B^{\mathrm{T}})[(\alpha A + \beta) * (\alpha B + \beta)^{\mathrm{T}}] - [A * (\alpha B + \beta)^{\mathrm{T}}] * [(\alpha A + \beta) * B^{\mathrm{T}}]\}^2 * 10^2}{\{(\alpha A + \beta) * B^{\mathrm{T}} + (\alpha A + \beta) * (\alpha B + \beta)^{\mathrm{T}}\} * \{A * B^{\mathrm{T}} + (\alpha A + \beta) * B^{\mathrm{T}}\}\{(A * B^{\mathrm{T}}) + A * (\alpha B + \beta)^{\mathrm{T}}\} * \{A * (\alpha B + \beta)^{\mathrm{T}} + (\alpha A + \beta) * (\alpha B + \beta)^{\mathrm{T}}\} + 10^{-4}}. \quad (1)$$

The GEP-CSI algorithm operated in a progressive mode as follows. SNPs obtained from the statistical analysis were organized and combined, and predominant SNP combinations were identified using the GEP algorithm. Then, the statistically significant SNPs were layered to form new combinations, with the primary SNP combinations being derived by further selection. This process was iterated until the fitness function was close to 100. In this study, the terms 'gene' and 'chromosome', pertain to codes in the algorithm symbolizing expressions, and are not related to the biological definitions of gene or chromosome.

2.9 | Protein-protein interaction (PPI) network

All gene names in the most correlated gene combinations were translated to the corresponding protein names via the UniProt website (https://www.uniprot.org/). The protein names were entered into the STRING website (https://string-db.org/) to extract PPI data for multiple proteins. The tab-separated values file and the interaction network diagram were both downloaded from the STRING website. The downloaded tab-separated values network file was then uploaded into Cytoscape software (v2.8.3, National Institute of General Medical Sciences) to visualize the PPI network.

3 | RESULTS

3.1 | Patients

The participants in this study were 30 Chinese patients with squamous cell nonsmall cell lung cancer who were treated with paclitaxel. The paclitaxel clearance data and the patient's genetic polymorphisms were split into a training data set and a validation data set. To ensure accurate outcomes, a similar proportion of patients with low clearance was maintained in the training and validation datasets. Before therapy, no statistically significant differences were detected between the high- and low-clearance groups for characteristics, such as age, sex, smoking history, drinking history, or tumor, node, and metastasis stage (Table 1). The median age of the entire cohort was 60 years (range 37–75 years). Most of the participants were male (87%), and most of them had a smoking history (83%). Furthermore, half (50%) of the entire cohort reported a history of alcohol consumption. The cancers of most of the participants were at the local metastasis stage, marked by lymph node metastases with no distant spread.

TABLE 2 Primary steps of the main program of the GEP-CSI data mining algorithm.

1. Initialize the start time of the program.
2. Initialize global variables ai and bi.
3. Initialize an empty list called fitness_all to store all fitness values.
4. Initialize an empty list called snp_all to store all snp column positions.
5. Generate a list of numbers and store it in read_snp.
6. Initialize an empty list called com_snp to store all possible combinations of two elements from read_snp.
7. For each combination in com_snp, perform the following steps:
   i. Set ai to the first element of the current combination.
   ii. positions.Set bi to the second element of the current combination.
   iii. Initialize a genome and set its functions and terminals.
   iv. Initialize a link genome and set its functions.
   v. Initialize an environment and set its population size, number of genes, number of homeotics, head length, homeotic head length, genome, and link genome.
   vi. Set the rates of various mutations in the environment.
   vii. Generate data and store it in inputsOutputs.
   viii. Run the genetic algorithm on inputsOutputs using the environment and evalFunction.
   ix. Store the highest fitness, mean fitness, and lowest fitness values in x, y_high, y_mean, and y_low respectively.
   x. Write the best fitness and snp column positions to a CSV file.
   xi. Print the current iteration number and the total number of iterations.
   xii. If the current fitness is greater than the previous best fitness, update the best fitness, ai, bi, chromosome, x, and y_high.
8. Print the best chromosome, best fitness, and snp column
9. Write the best fitness and snp column positions to a CSV file.
10. Plot the highest fitness values against the generation number.
11. Calculate the total time taken by the program and print it.

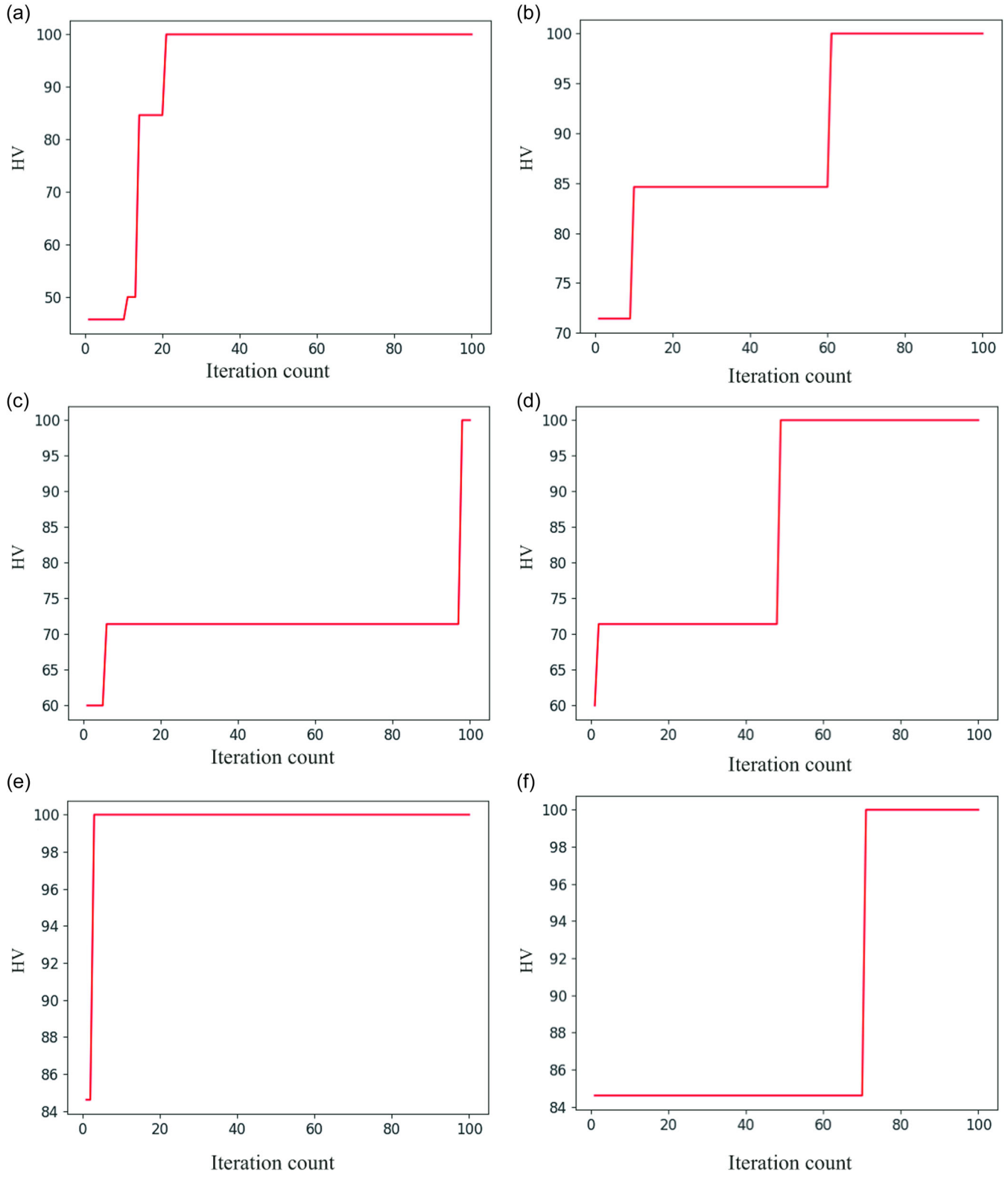

FIGURE 2 Evolutionary process of the GEP-CSI data mining algorithm iterations in three-gene combinations. (a) *ANKLE2*, *SLC22A31*, *LRRC2*. (b) *SUSD6*, *E4F1*, *ODC1*. (c) *SUSD6*, *MUC3A*, *C17orf97*. (d) *NUSAP1*, *SPATA2L*, *ACAN*. (e) *ANKLE2*, *SUSD6*, *NT5C3A*. (f) *FAM71A*, *SUSD6*, *NT5C3A*. HV, the highest value of fitness function.

3.2 | Pseudocode of the main program of the GEP-CSI algorithm

The program, which uses a genetic algorithm to optimize the fitness of a specified data set, was coded in Python. Every possible pair combination is generated systematically from a list of elements, and a genetic algorithm is applied to each pair to ascertain optimal fitness.

At its core, the code initializes a genome and establishes a population of 20 chromosomes. The genetic material in these chromosomes is represented as a series of symbols, where each symbol can be interpreted as a function, terminal, or other programmatic element. Genetic manipulations, such as mutation and crossing, alter the combinations of these symbols to produce new individuals in the population. The genetic algorithm is executed iteratively across 100 generations, producing data for each iteration. The fitness values (highest, lowest and average) are represented graphically for each generation. The algorithm presents the best-performing chromosome along with its fitness value and the associated values of the tuple's third element. The 'step' method is implemented by reordering and amalgamating the derived results with all the elements, subsequently leveraging the algorithm to deduce the best possible fitness value. An exhaustive description of the primary steps is provided in Table 2.

3.3 | Algorithm evolution

Genes with the top fitness values from the final population are replicated directly into the next generation, implying that genes with high fitness values are more likely to be inherited. Genetic evolution in a population is driven by various processes, including mutation, insertion element transposition, root insertion sequence transposition, gene transposition, one-point recombination, two-point recombination, and gene recombination. Expression evolution that mimics genetic processes hastens the algorithm's convergence, thereby reducing the number of required iterations. The evolutionary progression of the algorithm characterized by an improvement in fitness is illustrated in Figure 2.

3.4 | SNP-SNP interaction

A correlation analysis between SNP and clearance was conducted in R, yielding 239 SNPs with $p < 0.01$. The identified SNPs were paired and analyzed iteratively with the GEP-CSI algorithm to determine their association with clearance. A total of 28,441 SNP combinations with fitness function values of 14.81–99.99 were identified and recorded in a designated csv file. For algorithmic efficiency and comprehensive representation of biological effects, SNP combinations with fitness function values >70 were selected for incremental analysis. By integrating the selected combinations with the initial results and running the GEP-CSI algorithm, we obtained 202,911 combinations that had fitness values of 33.33–99.99. Among them, 9156 SNP combinations encompassing 177 genes had fitness function values of 99.99. We quantified the frequency of SNPs in the highest fitness combinations. The top 10 SNPs based on their number of occurrences and the corresponding chromosomal base positions and gene names are listed in Table 3.

TABLE 3 Top 10 SNPs according to their number of occurrences.

SNP ID	Number of occurrences	Gene	Chromosome	Base position
41	430	*MTCH2*	chr11	47644274
42	459	*MTCH2*	chr11	47644277
58	802	*RAB15*	chr14	65417070
109	492	*OR7A10*	chr19	14951898
142	604	*SFT2D3*	chr2	128459214
145	880	*UNC80*	chr2	210685100
146	894	*UNC80*	chr2	210685367
216	577	*NT5C3A*	chr7	33060946
220	532	*ARHGAP39*	chr8	145756170
233	574	*GSN*	chr9	124083614

3.5 | PPI network

The comprehensive PPI network encompasses 108 proteins produced by a single, protein-coding locus (Figure 3). The proteins are the nodes, and interactions between proteins are the edges. Alternative splicing or posttranslational modifications of these proteins were combined, and therefore each node symbolizes all the proteins that originated from a single protein-coding gene locus. The interactions suggest that the proteins collaborate toward shared functions, but are not necessarily physically bound to one another. We refined the initial PPI network by imposing a stricter filter with medium-confidence criteria. This resulted in a refined PPI network that included 22 relevant genes (Figure 4).

3.6 | Validation of the GEP-CSI algorithm mining results

With the GEP-CSI algorithm, we got gene expressions that contain logical strategies for SNP combinations. The identified SNP combinations and strategies were validated using the validation data set. Fitness function values were computed for the SNP combinations in the validation data set based on their respective strategies. The results showed

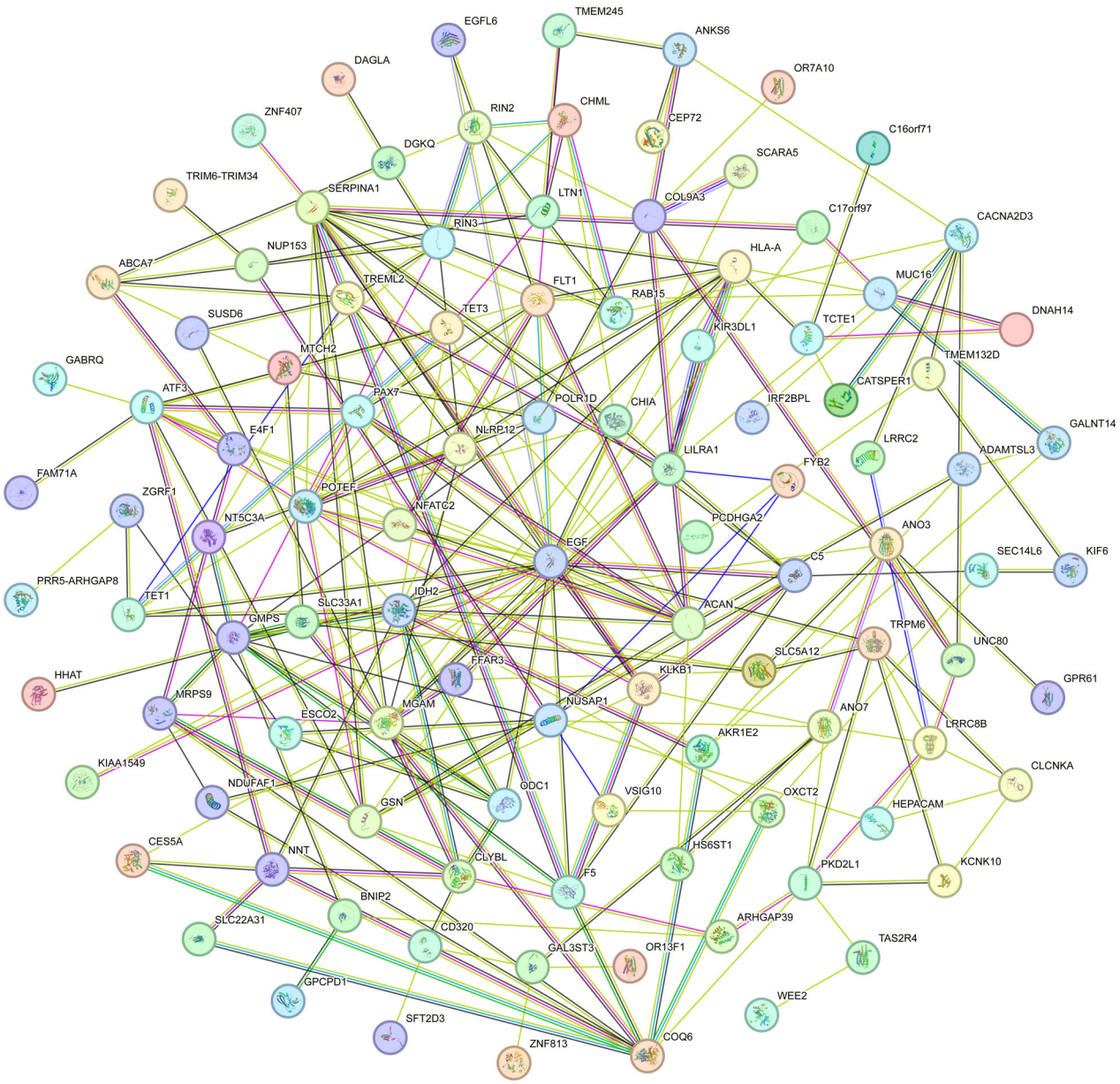

FIGURE 3 PPI network of the translated proteins of genes that had the highest fitness function values. Purple edges, known interactions (from curated databases); pink edges, known interactions (experimentally determined); green, red, and blue edges, predicted interactions.

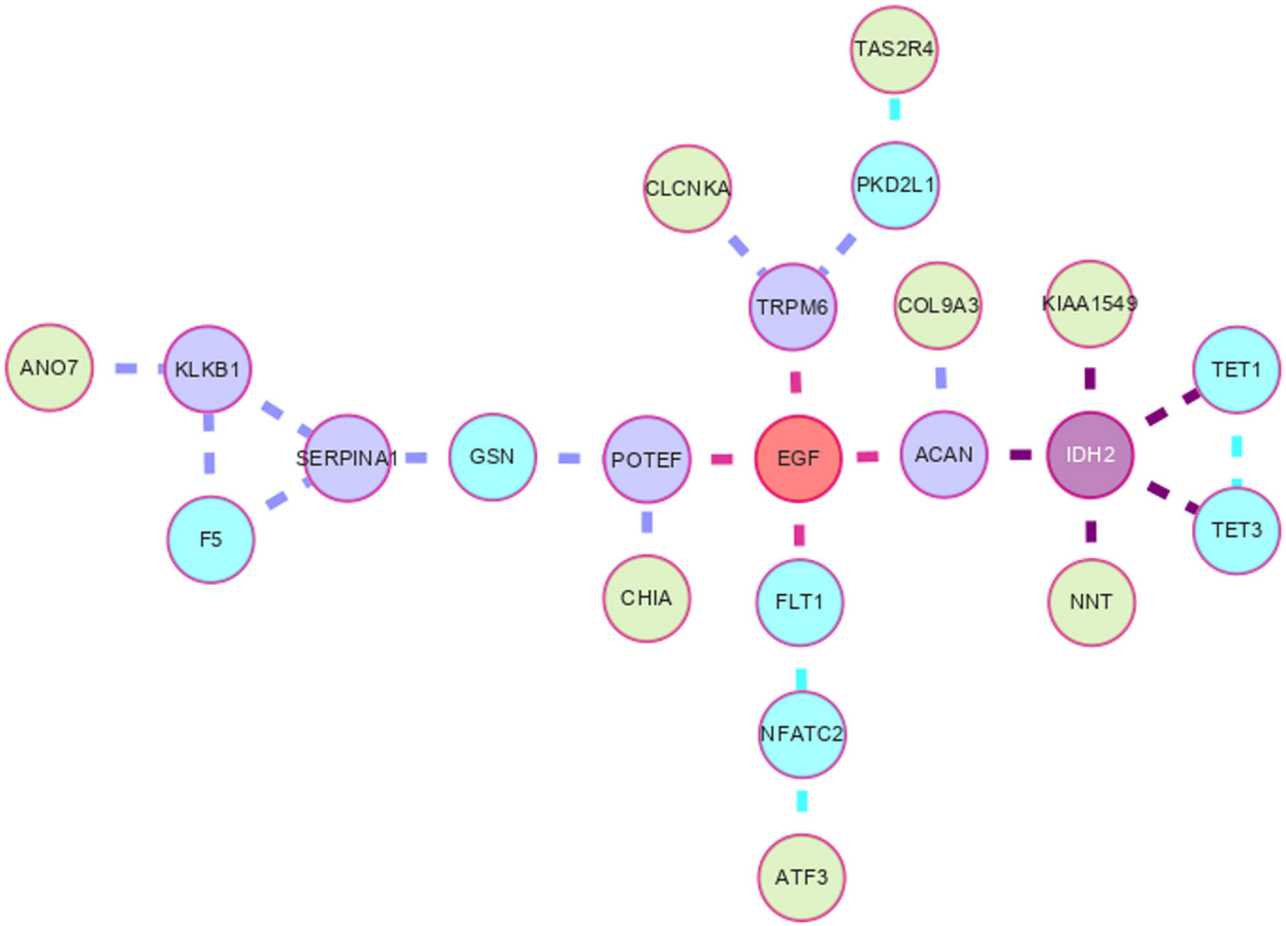

FIGURE 4 Refined PPI network using a stricter filter with medium-confidence criteria (0.400). Purple node, five interactions; red node, four interactions; dark blue node, three interactions; blue node, two interactions; green node, one interaction.

that 1184 combinations and strategies achieved the highest fitness function value of 99.99. Conversely, when SNPS are randomly combined, there is a one in two [6] chance that the calculated result is the highest fitness function value, which is substantially lower than the value achieved using our mined results. These findings suggest there is a robust correlation between the mined SNP combinations, strategies, and paclitaxel clearance, and also underscore the efficacy of the GEP-CSI data mining algorithm. We posited that augmenting the data set may further enhance the predictive accuracy of the algorithm for paclitaxel clearance.

4 | DISCUSSION

4.1 | Relevant SNPs

Paclitaxel is metabolized primarily in the liver, and the predominant elimination route for both the parent drug and its hydroxylated metabolites is biliary excretion into the feces. Previously identified genes involved in paclitaxel metabolism and transport include *OATP1B3* (also known as *SLCO1B3*), *OAT2* (also known as *SLC22A7*), *NR1I2* [20], *CYP1B1*, *CYP2C8* [21], *CYP3A4* [22], *ABCB1*, *ABCC1*, *ABCC2*, and *ABCG2*. Surprisingly, the genes that were previously reported to significantly influence the variability of paclitaxel clearance were absent in our final data mining results. This discrepancy may be because the algorithm focused predominantly on the effect of SNP combinations, rather than individual genes, on paclitaxel clearance. Furthermore, the impact of gene expression on paclitaxel pharmacokinetics is a complex nonlinear process that may be influenced by different drug formulations. This nuanced combinatorial effect becomes even more apparent when the translated proteins of previously reported genes are included in the PPI network (Figure 5). This finding suggests that genes from multiple networks may indirectly influence other genes that are pivotal to paclitaxel clearance. For instance, *SERPINA1* interacted with *CYP3A4*, *ATF3* interacted with *CYP1B1*, and *ABCG2* and *ABCB1* both interacted with epidermal growth factor (EGF) in the complete PPI network (Figure 5).

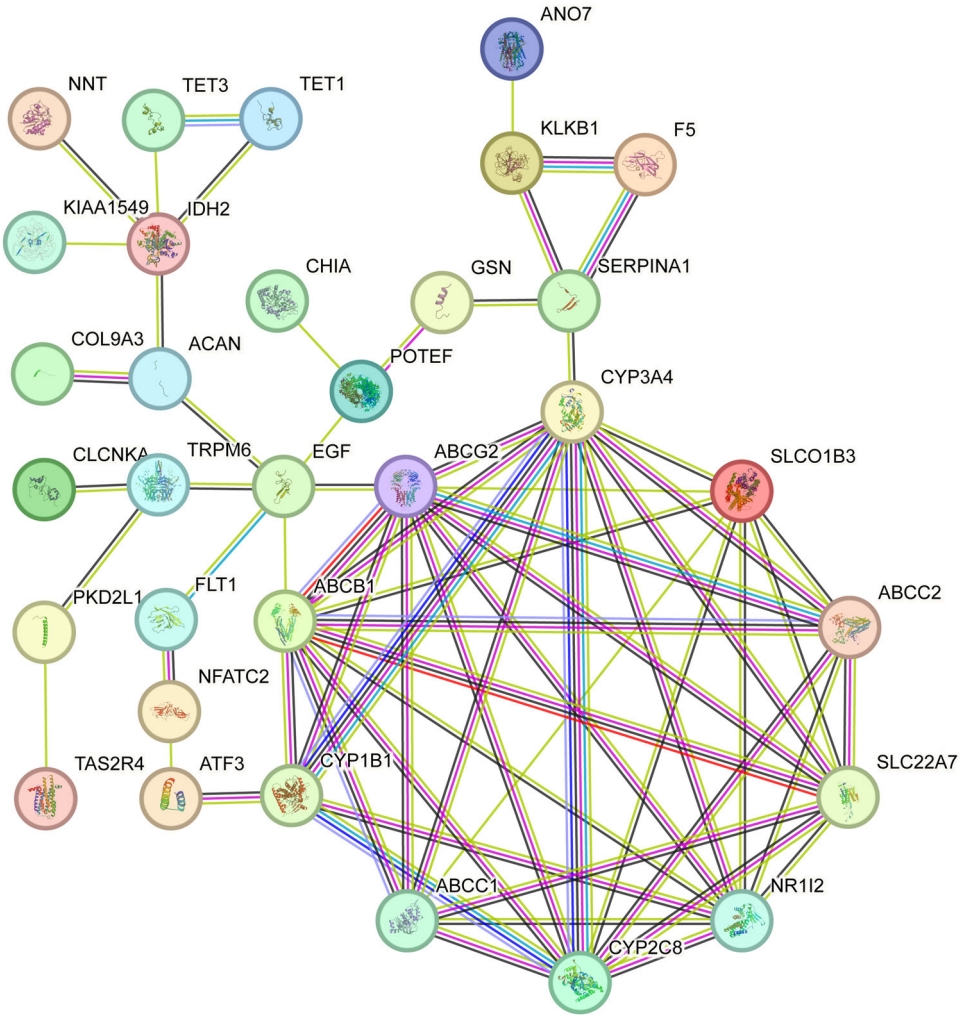

FIGURE 5 PPI network with the translated proteins of previously reported genes included. Purple edges, known interactions (from curated databases); pink edges, known interactions (experimentally determined); green, red, and blue edges, predicted interactions.

4.2 | SNPs matching STRING interactions

A total of 251 interactions were contained in the complete PPI network. By jointly analyzing the results for the three-SNP combinations and cross-referencing them with the gene interaction data on the STRING website, we discerned three patterns. First, we found two distinct results that involved three specific gene combinations in which all the genes interacted with one another. Second, we identified 26 results in which a single gene served as a bridge that connected the three genes and mediated interactions. These interactions encompassed 22 unique gene combinations and 29 individual genes. Lastly, we found 669 results in which only two genes interacted, encompassing 90 distinct gene combinations and 75 individual genes. The genes involved in the pairwise interactions, such as *FLT1*, *EGF*, and *MUC16*, and so forth, have stronger associations and are more likely to affect the clearance of paclitaxel.

4.3 | FLT1, EGF, and MUC16 interactions

The GEP-CSI data mining algorithm identified an interaction between *EGF* and *MUC16*; however, this interaction was not evident in the medium-confidence PPI network. Even so, the interaction between MUC16 and EGF receptors (EGFR) has been reported to enhance the activation of carcinogenic signaling pathways, especially when MUC16 undergoes aberrant glycosylation [23]. Moreover, the association between EGFR and FLT1 was confirmed to correlate with the risk of sentinel lymph node metastasis and/or patient mortality, underscoring the link between these genes [24].

4.4 | Cytochrome P450 enzymes

EGF has been shown to have vital properties that counteract and reverse the effects of TGF-β1 [25]. This antagonistic role prompts a cascade of cellular

alterations, notably inhibiting pathways involved in cell cycle progression, DNA replication, and in particular, the metabolism of xenobiotics by cytochrome P450 [26]. Additionally, during the in vitro development of goat follicles, EGF had various impacts on the mRNA levels of EGF, EGF-R, FSR-R, and P450 aromatase depending on the stage (early or late) of follicular development [27].

4.5 | Clinical practice

For patients undergoing paclitaxel chemotherapy, it is advisable to prioritize SNPs associated with genes in high-confidence networks together with SNPs linked to previously reported genes. The gene expression data obtained using the GEP-CSI algorithm can be transformed into conventional expression values. When combined with the results of clinical SNP tests, the expression data can be a valuable tool for predicting paclitaxel clearance in humans. The predictive information can serve as a reference for healthcare professionals to make necessary adjustments to medication dosages.

This study has some potential limitations. First, the sample size was small. To mitigate this limitation, we validated the finding that mining gene expressions significantly enhanced the accuracy of the prediction results on an independent data set, thereby demonstrating the effectiveness of the GEP-CSI data mining algorithm. Second, the interference of other covariates in the context of clinical trials is another limitation. In previous studies, population pharmacokinetic methods have been used to estimate the impact of covariates on paclitaxel clearance rates [19]. In this study, we conducted a statistical analysis of intergroup differences in key variables and found no significant disparities.

In future investigations, molecular biology techniques will be used to delve deeper into the mechanisms of the SNP interactions identified in this study. Additionally, the GEP-CSI data mining algorithm will be applied to uncover SNP interactions across diverse fields of research.

5 | CONCLUSION

A new algorithm for uncovering SNP interactions was described and was shown to provide insights into the prediction of nonlinear relationships in biological outcomes. However, because our sample size was small, further research is essential to validate the relationship between these SNPs and paclitaxel clearance, and to elucidate the underlying molecular mechanisms.

AUTHOR CONTRIBUTIONS

Wei Chen: Data curation (equal); investigation (lead); methodology (lead); project administration (lead); validation (equal); writing—original draft (lead). **Haiyan Zhou**: Formal analysis (equal); investigation (equal); resources (equal). **Mingyu Zhang**: Data curation (equal); investigation (equal). **Yafei Shi**: Software (equal); visualization (equal). **Taifeng Li**: Investigation (equal); software (equal). **Di Qian**: Investigation (equal). **Jun Yang**: Conceptualization (equal); investigation (equal); resources (equal); writing—review and editing (equal). **Feng Yu**: Conceptualization (equal); investigation (equal); resources (equal); supervision (equal); writing—review and editing (equal). **Guohui Li**: Conceptualization (equal); funding acquisition (equal); resources (equal); writing—review and editing (equal).

ACKNOWLEDGMENTS

None.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT

The study protocol was approved by the Institutional Review Board of the Cancer Hospital, Chinese Academy of Medical Sciences (Approval number: 15-123/1050), and it was compliant with the Helsinki Declaration of 1975, as revised in 2008.

INFORMED CONSENT

All patients provided written informed consent at the time of entering this study.

ORCID

Wei Chen http://orcid.org/0000-0002-4015-6021

REFERENCES

1. Young KL, Graff M, North KE, Richardson AS, Bradfield JP, Grant SFA, et al. Influence of SNP*SNP interaction on BMI in European American adolescents: findings from The National Longitudinal Study of Adolescent Health. Pediatr Obes. 2016;11(2):95–101. https://doi.org/10.1111/ijpo.12026
2. Lee S, Kwon MS, Oh JM, Park T. Gene-gene interaction analysis for the survival phenotype based on the Cox model. Bioinformatics. 2012;28(18):i582–8. https://doi.org/10.1093/bioinformatics/bts415
3. Gui J, Moore JH, Kelsey KT, Marsit CJ, Karagas MR, Andrew AS. A novel survival multifactor dimensionality

reduction method for detecting gene–gene interactions with application to bladder cancer prognosis. Hum Genet. 2011;129(1):101–10. https://doi.org/10.1007/s00439-010-0905-5
4. Wagner M, Tupikowski K, Jasek M, Tomkiewicz A, Witkowicz A, Ptaszkowski K, et al. SNP-SNP interaction in genes encoding PD-1/PD-L1 axis as a potential risk factor for clear cell renal cell carcinoma. Cancers. 2020;12(12):3521. https://doi.org/10.3390/cancers12123521
5. Liu L, Zhai W, Wang F, Yu L, Zhou F, Xiang Y, et al. Using machine learning to identify gene interaction networks associated with breast cancer. BMC Cancer. 2022;22(1):1070. https://doi.org/10.1186/s12885-022-10170-w
6. Moore R, Ashby K, Liao TJ, Chen M. Machine learning to identify interaction of single-nucleotide polymorphisms as a risk factor for chronic drug-induced liver injury. Int J Environ Res Public Health. 2021;18(20):10603. https://doi.org/10.3390/ijerph182010603
7. Iiyama S, Fukaya K, Yamaguchi Y, Watanabe A, Yamamoto H, Mochizuki S, et al. Total synthesis of paclitaxel. Org Lett. 2022;24(1):202–6. https://doi.org/10.1021/acs.orglett.1c03851
8. Abu Samaan TM, Samec M, Liskova A, Kubatka P, Büsselberg D. Paclitaxel's mechanistic and clinical effects on breast cancer. Biomolecules. 2019;9(12):789. https://doi.org/10.3390/biom9120789
9. Nakao M, Fujita K, Suzuki Y, Arakawa S, Sakai Y, Sato H, et al. Nab-paclitaxel monotherapy for relapsed small cell lung cancer: retrospective analysis and review. Anticancer Res. 2020;40(3):1579–85. https://doi.org/10.21873/anticanres.14105
10. Blair HA, Deeks ED. Albumin-bound paclitaxel: a review in non-small cell lung cancer. Drugs. 2015;75(17):2017–24. https://doi.org/10.1007/s40265-015-0484-9
11. Kampan NC, Madondo MT, McNally OM, Quinn M, Plebanski M. Paclitaxel and its evolving role in the management of ovarian cancer. BioMed Res Int. 2015;2015:1–21. https://doi.org/10.1155/2015/413076
12. Rodríguez-Antona C. Pharmacogenomics of paclitaxel. Pharmacogenomics. 2010;11(5):621–3. https://doi.org/10.2217/pgs.10.32
13. Al-Mahayri ZN, AlAhmad MM, Ali BR. Current opinion on the pharmacogenomics of paclitaxel-induced toxicity. Expert Opin Drug Metab Toxicol. 2021;17(7):785–801. https://doi.org/10.1080/17425255.2021.1943358
14. Yang YH, Mao JW, Tan XL. Research progress on the source, production, and anti-cancer mechanisms of paclitaxel. Chin J Nat Med. 2020;18(12):890–7. https://doi.org/10.1016/S1875-5364(20)60032-2
15. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010;42(11):937–48. https://doi.org/10.1038/ng.686
16. Stage TB, Bergmann TK, Kroetz DL. Clinical pharmacokinetics of paclitaxel monotherapy: an updated literature review. Clin Pharmacokinet. 2018;57(1):7–19. https://doi.org/10.1007/s40262-017-0563-z
17. De Graan AJM, Elens L, Smid M, Martens JW, Sparreboom A, Nieuweboer AJM, et al. A pharmacogenetic predictive model for paclitaxel clearance based on the DMET platform. Clin Cancer Res. 2013;19(18):5210–7. https://doi.org/10.1158/1078-0432.CCR-13-0487
18. Yang Z, Wen Y, Chen Y. sEMG-based drawing trace reconstruction: a novel hybrid algorithm fusing gene expression programming into Kalman filter. Sensors. 2018;18(10):3296. https://doi.org/10.3390/s18103296
19. Zhou H, Yan J, Chen W, Yang J, Liu M, Zhang Y, et al. Population pharmacokinetics and exposure-safety relationship of paclitaxel liposome in patients with non-small cell lung cancer. Front Oncol. 2021;10:1731. https://doi.org/10.3389/fonc.2020.01731
20. Harmsen S, Meijerman I, Beijnen JH, Schellens JHM. Nuclear receptor mediated induction of cytochrome P450 3A4 by anticancer drugs: a key role for the pregnane X receptor. Cancer Chemother Pharmacol. 2009;64(1):35–43. https://doi.org/10.1007/s00280-008-0842-3
21. Dai D, Zeldin DC, Blaisdell JA, Chanas B, Coulter SJ, Ghanayem BI, et al. Polymorphisms in human CYP2C8 decrease metabolism of the anticancer drug paclitaxel and arachidonic acid. Pharmacogenetics. 2001;11(7):597–607. https://doi.org/10.1097/00008571-200110000-00006
22. Cresteil T, Monsarrat B, Dubois J, Sonnier M, Alvinerie P, Gueritte F. Regioselective metabolism of taxoids by human CYP3A4 and 2C8:structure-activity relationship. Drug Metab Dispos. 2002;30(4):438–45. https://doi.org/10.1124/dmd.30.4.438
23. Thomas D, Sagar S, Liu X, Lee HR, Grunkemeyer JA, Grandgenett PM, et al. Isoforms of MUC16 activate oncogenic signaling through EGF receptors to enhance the progression of pancreatic cancer. Mol Ther. 2021;29(4):1557–71. https://doi.org/10.1016/j.ymthe.2020.12.029
24. Park JY, Amankwah EK, Anic GM, Lin HY, Walls B, Park H, et al. Gene variants in angiogenesis and lymphangiogenesis and cutaneous melanoma progression. Cancer Epidemiol Biomarkers Prevent. 2013;22(5):827–34. https://doi.org/10.1158/1055-9965.EPI-12-1129
25. Wang P, Yang AT, Cong M, Liu TH, Zhang D, Huang J, et al. EGF suppresses the initiation and drives the reversion of TGF-β1-induced transition in hepatic oval cells showing the plasticity of progenitor cells. J Cell Physiol. 2015;230(10):2362–70. https://doi.org/10.1002/jcp.24962
26. Wang P, Cong M, Liu T, Yang A, Sun G, Zhang D, et al. The characteristics variation of hepatic progenitors after TGF-β1-induced transition and EGF-induced reversion. Stem Cells Int. 2016;2016:1–10. https://doi.org/10.1155/2016/6304385
27. Silva CMG, Castro SV, Faustino LR, Rodrigues GQ, Brito IR, Rossetto R, et al. The effects of epidermal growth factor (EGF) on the in vitro development of isolated goat secondary follicles and the relative mRNA expression of EGF, EGF-R, FSH-R and P450 aromatase in cultured follicles. Res Vet Sci. 2013;94(3):453–61. https://doi.org/10.1016/j.rvsc.2012.12.002

How to cite this article: Chen W, Zhou H, Zhang M, Shi Y, Li T, Qian D, et al. Novel progressive deep learning algorithm for uncovering multiple single nucleotide polymorphism interactions to predict paclitaxel clearance in patients with nonsmall cell lung cancer. Cancer Innov. 2024;3:e110. https://doi.org/10.1002/cai2.110