International Journal of
Environmental Research and Public Health

MDPI

Article

Machine Learning to Identify Interaction of Single-Nucleotide Polymorphisms as a Risk Factor for Chronic Drug-Induced Liver Injury

Roland Moore, Kristin Ashby, Tsung-Jen Liao and Minjun Chen *

Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Rd, Jefferson, AR 72079, USA; rolmoore2000@gmail.com (R.M.); kristin.mceuen@fda.hhs.gov (K.A.); Tsung-Jen.Liao@fda.hhs.gov (T.-J.L.)
* Correspondence: Minjun.chen@fda.hhs.gov; Fax: +1-870-543-7865

Abstract: Drug-induced liver injury (DILI) is a major cause of drug development failure and drug withdrawal from the market after approval. The identification of human risk factors associated with susceptibility to DILI is of paramount importance. Increasing evidence suggests that genetic variants may lead to inter-individual differences in drug response; however, individual single-nucleotide polymorphisms (SNPs) usually have limited power to predict human phenotypes such as DILI. In this study, we aim to identify appropriate statistical methods to investigate gene–gene and/or gene–environment interactions that impact DILI susceptibility. Three machine learning approaches, including Multivariate Adaptive Regression Splines (MARS), Multifactor Dimensionality Reduction (MDR), and logistic regression, were used. The simulation study suggested that all three methods were robust and could identify the known SNP–SNP interaction when up to 4% of genotypes were randomly permutated. When applied to a real-life DILI chronicity dataset, both MARS and MDR, but not logistic regression, identified combined genetic variants having better associations with DILI chronicity in comparison to the use of individual SNPs. Furthermore, a simple decision tree model using the SNPs identified by MARS and MDR was developed to predict DILI chronicity, with fair performance. Our study suggests that machine learning approaches may help identify gene–gene interactions as potential risk factors for better assessing complicated diseases such as DILI chronicity.

Keywords: drug-induced liver injury; chronicity; machine learning; SNP; genotype; gene–gene interactions; epistasis; splines

Citation: Moore, R.; Ashby, K.; Liao, T.-J.; 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*, 10603. https://doi.org/10.3390/ijerph182010603

Academic Editor: Matteo Goldoni

Received: 9 August 2021
Accepted: 5 October 2021
Published: 10 October 2021

1. Introduction

As the primary organ responsible for metabolism, the liver is vulnerable to injury caused by drugs and drug metabolites [1,2]. Drug-induced liver injury (DILI) is one of the main reasons for halting drug development processes [3], and has caused over 50 approved drugs to be withdrawn from the market [4,5]. Moreover, even when a drug is deemed safe and approved for public use, a relatively small fraction of the population taking the drug may experience liver damage, and therefore identifying risk factors is of particular importance for preventing DILI [6,7]. The risk factors for DILI include age, sex, drug properties, and genetic variations [8,9].

Genetic factors have attracted increasing attention, and recent studies have shown that genetic variants in drug-metabolizing enzymes and human leukocyte antigen (HLA) were associated with the occurrence of DILI [10–12]. Many efforts have focused on the genetic risk factors of DILI, and several genetic variants such as single-nucleotide polymorphism (SNP) and insertion/deletion have been identified as risk factors associated with the use of specific drugs [13–16]. For example, carriers of HLA-B*57:01 are 80 times more likely to develop flucloxacillin-induced DILI than those who do not carry the variant [17]. The association of liver injury caused by specific drugs or groups of drugs with polymorphisms

in HLA was also reported [18,19]. However, most of these studies investigated the association of each SNP individually [20,21]; SNP–SNP interaction as a potential risk factor for drug-induced liver injury was seldom reported.

Pairwise SNP–SNP interactions have been recognized to play a role in assessing disease susceptibility and drug responses [22,23]. Several statistical methods have been used to identify SNP–SNP interactions [24]. For example, Cook et al. [25] employed classification and regression trees and multivariate adaptive regression spline (MARS) models to explore the presence of genetic interactions for ischemic stroke. Moore and Williams [26] successfully applied the multifactor dimensionality reduction (MDR) method to identify gene–gene interactions in essential hypertension. Notably, SNPs themselves come in large numbers, sometimes in the hundreds of thousands; including their interactions, the number of factors which need to be considered could expand exponentially. The issue of dimensionality becomes a challenge of statistical genetic data analysis, especially when the sample size is relatively small compared to the number of factors. Certain SNPs tend to occur together and are not statistically independent, which can complicate analysis. The combination of SNPs that are preferably uncorrelated to each other may improve the classification performance [27]. Statistical genetic data analysis is generally computationally intensive and time-consuming, and classical statistical methods such as regression are less feasible for use with such high-dimensional data. Thus, appropriate statistical/machine learning methodologies are critical for overcoming the challenges of analyzing gene–gene interactions.

In this study, we aim to investigate SNP–SNP interaction as a potential risk factor for predicting DILI chronicity using machine learning approaches. We briefly introduce three machine learning methods, including MARS, MDR, and logistic regression. Then, we use simulated data with known pairwise SNP–SNP interactions to evaluate the accuracies and robustness of these methods. Next, we apply the three methods to identify SNP interactions associated with chronic DILI. Finally, we evaluate the predictive performance of the identified SNPs by using a simple decision tree model to assess whether the presence of these SNP–SNP interactions is associated with an increased risk of DILI chronicity.

2. Materials and Methods

2.1. DILI Chronicity Cohort

Cases in this study were collected by International Serious Adverse Event Consortium (iSAEC), a large international collaborative study, including partnerships with the United Kingdom, Sweden, Spain, Germany, France, Switzerland, the Netherlands, Australia, and Finland. The study protocols were approved by local ethics committees, and the informed consent was obtained from all subjects involved in the study. Inclusion criteria for DILI cases followed the clinical chemistry criteria defined [28]. Specifically, a case must have either alanine aminotransferase elevated $\geq 5\times$ the upper level of normal, or elevated alkaline phosphatase $\geq 2\times$ the upper level of normal, or elevated levels of alanine aminotransferase $\geq 3\times$ the upper level of normal while bilirubin concentrations are also two-fold higher than the upper level of normal. Causality assessment was conducted using the Roussel Uclaf Causality Assessment Method scoring system and expert review, consisting of a panel of three hepatologists. Only cases with a causality scale of probable with a score of greater than or equal to six were included. Cases with preexisting liver disease were excluded. DILI chronicity was determined by the time of patient recovery, which was defined as the days from the medication stop date until the date that the patient's serum liver biochemistries returned to normal. For this dataset, chronic DILI is defined as liver injury without recovery for six months or more, termed chronic; acute DILI is defined as liver injury with recovery within six months, termed acute.

A dataset with 271 patients (33 chronic versus 238 acute DILI) were used to assess the three machine learning methods for identifying SNP–SNP interaction as a risk factor for chronic DILI. DNA preparation is described here [29] and standard quality control procedures were followed [30]. Samples were genotyped using the HumanOmniExpressExome-

8v1 or the Illumina HumanCoreExome-12 v1.0 BeadChip, from which 872 SNPs associated with the genes in bile acid pathways were retrieved for analysis. Bile acid pathways were downloaded from the Molecular Signature Database (MSigDB) [31] and included pathways involved in bile acid synthesis, recycling, and transport.

2.2. Machine Learning Approaches to Identify SNP–SNP Interactions

Assuming Y is the phenotype, such as chronic DILI (binary; chronic or acute); $P(Y = 1) = p$, A is an SNP, B is another SNP, and $A \times B$ is an SNP–SNP interaction factor, the parametric model for the logistic regression is as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 A \times B \quad (1)$$

by contrast, the non-parametric statistical method model is:

$$Y = f(A, B) \quad (2)$$

The statistical methodologies for identifying gene–gene interactions were reviewed thoroughly [22,23]. Here, three frequently used methods including MARS [32], MDR [33,34], and logistic regression [35] were selected for evaluation based on availability and use.

2.2.1. MARS Approach

The MARS approach is an adaptive algorithm for identifying SNP–SNP interactions, which fits models using flexible regression modeling, including non-linear components and interactions [32]. It was reported to perform better than typical logistic regression methods when applied to the discovery of interactions among genes without strong marginal effects [25].

In MARS, the entire dataset was divided into multiple smaller regression subsets through automatic dataset-determined knots. Each of these subsets comes with a basis function (usually a linear spline function), and all significant basis functions are aggregated to obtain the overall regression model. MARS gives a regression-like output model with certain basis functions of the predictors. It builds models of the following equation form:

$$\hat{f} = \sum_{i=1}^{k} C_i B_i(x) \quad (3)$$

where each C_i is the coefficient multiplying a basis function $B_i(x)$, k is the number of knots, and $\hat{f}$ is the model estimation.

The regression model of MARS is derived directly from the data, through a data-driven automatic set of basis functions with their corresponding coefficients. These basis functions are derived based on automatic data-driven hinge values in the data, also referred to as knots. As a specific linear relationship between the response and predictors in a data subset is being modeled, the knot automatically identifies the point of direction change from that relationship. This point of change becomes a starting point for a new subset of another relationship, and the process continues to the end with new knots and relationships. In this way, the model captures both linear and non-linear relationships, with corresponding interactions.

The MARS model is built using a forward-and-backward selection process in the following way: the forward selection first combines all significant basis regression functions, and their interactions, and then backward selection prevents overfitting by pruning the result from the forward selection, removing basis functions one at a time, and selecting the model with the lowest Generalized-Cross-Validation (GCV) score. This process will yield the optimized final model. The GCV equation is written as follows:

$$GCV = \frac{\sum_{i=1}^{N} (y_i - \hat{f})^2}{(1 - C/N)^2} \quad (4)$$

where $C = 1 + p * k$ with k as independent number of base functions and p as the penalty for adding a base function. Here, y_i ($i = 1$ to N) is the value of phenotype for the observation i, and N is number of observations, and $\hat{f}$ is from the Equation (3).

2.2.2. MDR Approach

The MDR approach is a non-parametric dimension-reducing method that was created primarily to detect gene–gene and/or gene–environment interactions [33,34,36]. Compared with MARS and other statistical methods which usually handle interactions between two factors, MDR has the power to identify statistically significant high-order interactions among three or more factors [33]. Its central concept is to utilize a specially designed strategy to transform multilocus information to a one-dimensional model. In this strategy, a subset of n genetic factors are first selected, and then these n-genetic factors and their possible multifactor classes are represented by n-dimensional space, in which each multifactor cell will be labeled as high-risk or low-risk group. In this way, the case-control model for multilocus genotypes will be converted into classifications of highrisk and low-risk, which reduces the n-dimensional model to a one-dimensional model. In the next step, the prediction error estimated through cross-validation is used to evaluate the selected n-genetic factor. For each n-genetic factor combination, a single model that minimizes the average classification error in the cross-validation training sets is selected. This will result in a list of best models, one for each value of n-genetic factor. Among these classification models, the combination of the genetic factors and the model they built that minimizes the average prediction error across the prediction errors in the cross-validation testing sets will be selected as the final one.

2.2.3. Logistic Regression Approach

Logistic regression is the mostly used method for identifying gene–gene interaction; however, it is not so successful in its handling of the datasets with a large number of SNPs to consider, and weak or no marginal effects of SNPs [37]. Here, Random Forest was used together with logistic regression approach [38]. This approach used multiple decision trees as a learning ensemble to give a single output. Each tree acts on random bootstrap subsets of the data, as well as random subsets of the SNP predictors to make predictions. Each predictor represents the node of a tree, and a route links a sequence of predictors from the roots to the leaves. The predictions from each of these trees are then aggregated to give the overall output prediction. Majority voting is employed in classification. Random Forest produces a table ranking of the SNP predictors in the order of importance of their contributions to the phenotype. Logistic regression is then applied to the first few variables to determine the interacting SNPs that are linked to the phenotype.

2.3. Data Analysis

Here, we firstly employed a simulated dataset with known SNP–SNP interactions to evaluate the performance and robustness of these machine learning methods. The simulated dataset was modified from the dataset retrieved from the MDR software package. The dataset contains 250 observations, including 125 positive and negative phenotype responses, respectively, and 25 SNPs, of which SNP4 and SNP9 are a pair of known SNP interaction. The frequencies of the minority genotypes of SNP4 and SNP9 are 21% (52/250) and 23% (57/250). To test the robustness of the methodologies, we permutated the genotypes of 23 SNPs, except SNP4 and SNP9. In the first test, two of the 250 observations were selected for permutation, while in the second test 10 observations were permutated. Each analysis includes 100 permutations.

We also applied these machine learning approaches to a real-life DILI chronicity dataset. Figure 1 briefly illustrated the pipeline for identifying SNP–SNP interactions as potential risk factors which are associated with chronic drug-induced liver injury. Firstly, based on the dataset of 271 patients and 872 SNPs, three machine learning methods including MARS, MDR and Random Forest plus logistic regression were utilized to identify

SNP–SNP interactions linked to DILI chronicity. Next, only the SNP–SNP interactions have the better associations with DILI chronicity, which is measured by odds ratio, than the individual SNPs were considered. Finally, The SNPs from the selected interactions were pooled together as candidate predictors, and then a decision tree model using classification and regression trees (CART) algorithm is developed.

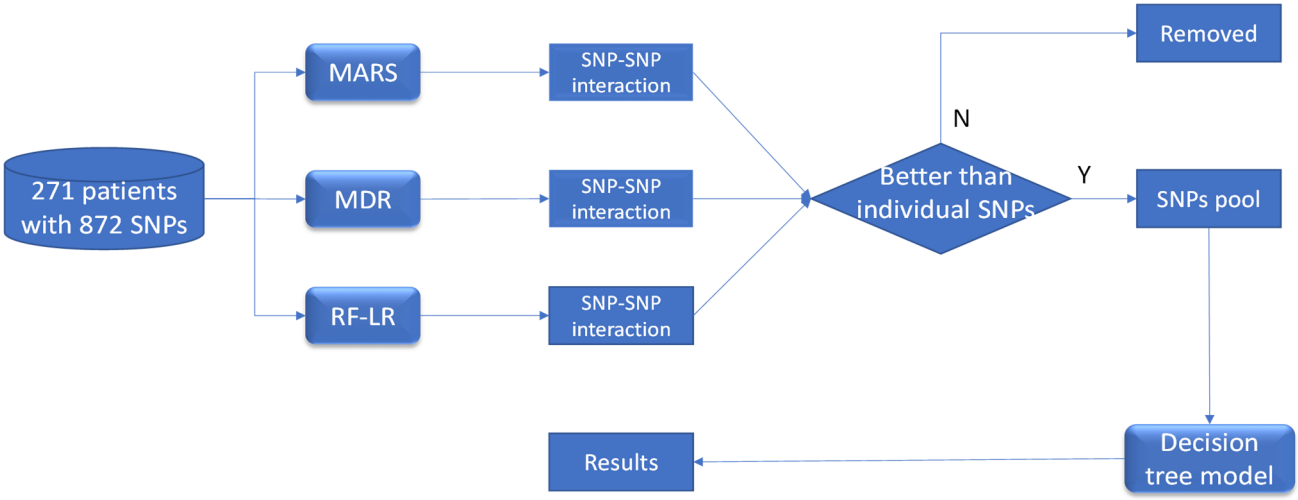

Figure 1. The diagram of the working flow to identify SNP–SNP interaction as a potential risk factor for chronic drug-induced liver injury. Specifically, multivariate adaptive regression spline (MARS), multifactor dimensionality reduction (MDR) and Random Forest plus logistic regression (RF-LR) were used to identify the SNP–SNP interaction terms linked to chronic DILI. Only the SNP–SNP interactions which had better association with DILI chronicity than individual SNPs were kept. All these SNPs of the selected interactions were pooled together as the candidate predictors, and a decision tree model is developed using classification and regression trees algorithm.

All analyses except those specifically mentioned were performed using R (version 3.6.1) [39] and the MDR package [40] for multifactor dimensionality reduction approach, the randomForest package [41] for Random Forest algorithm, the Rpart package [42] for decision tree model, and the Stats package for logistic regression and Fisher exact test. Multivariate adaptive regression spline approach was implemented using the MARS engine of Salford Predictive Modeler 8.0 from MinTab [43].

3. Results

3.1. Simulation Analysis

Three machine learning methods were evaluated using the simulated dataset with a known SNP–SNP interaction (SNP4 + SNP9). As shown in Table 1, all three methods successfully identified the known SNP interaction. We further conducted a permutation analysis to test the robustness of the three methods. In the first test, the genotypes of two observations randomly selected from 250 observations were permutated. As a result, all three methods successfully identified the known SNP interaction. Even when the number of permutated observations was increased from 2 to 10 of 250 observations, MDR slightly decreased in recovery rate and correctly identified the known SNP–SNP interactions among 97% of 100 permutations, while logistic regression and MARS still maintained a 100% recovery rate of the known SNP–SNP interaction. In other words, these machine learning approaches were robust when up to 4% (or 10 of 250) of the observations were permutated.

Table 1. Recovery rate of the known SNP4–SNP9 interaction by using three machine learning approaches in the original simulation dataset and permutation tests.

	Recovery Rate of the Known SNPs Interaction		
	MARS *	**MDR ***	**RF-LR ***
Original simulation dataset	100%	100%	100%
Permutation test 1 (2 of 250 observations permutated)	100% (100/100)	100% (100/100)	100% (100/100)
Permutation test 2 (10 of 250 observations permutated)	100% (100/100)	97% (97/100)	100% (100/100)

* MARS: Multivariate Adaptive Regression Splines; MDR: Multifactor Dimensionality Reduction; RF-LR: Random Forest plus Logistic Regression.

3.2. Chronic DILI Data Analysis

We first checked the association of DILI chronicity with individual SNPs using the Fisher exact test. Five individual SNPs from SNP–SNP pairs selected by the machine learning methods below were tested. As shown in Table 2, two SNPs (rs6487213 and rs5417) were significantly associated with DILI chronicity, with odds ratios of 3.28 (95% CI: 1.57–7.09, $p = 0.002$) and 3.01 (95% CI: 1.44–6.43, $p = 0.004$), respectively. The other three cited SNPs were not statistically significant.

We also applied the three machine learning methods to identify the SNP interaction terms that were linked to DILI chronicity. As shown in Table 2, MARS identified an interaction of a two-SNP combination (rs6487213 + rs3785157), while RF-LR detected another two-SNP interaction (rs5417 + rs3785157), and MDR found a three-SNP interaction (rs5417 + rs7658048 + rs12453290). Importantly, the SNP interactions identified by MARS and MDR have an improved association between the presence of genotypes and the chronicity of DILI. The odds ratio of the SNP interaction rs6487213 + rs3785157 identified by the MARS approach increased to 4.74 (95% CI: 2.14–10.39, $p < 0.001$), while the odds ratios of rs6487213 and rs3785157 alone are 3.28 (95% CI: 1.57–7.09, $p = 0.002$) and 1.49 (0.72–3.14, $p = 0.282$), respectively. Additionally, 30.4% of the combined genotypes of CC + CC by rs6487213 + rs3785157 were reported with DILI chronicity, as compared to 20.8% of the genotype CC of rs6487213 alone. The MDR approach identified a three-SNP interaction, and its odds ratio of 4.19 (95% CI: 1.36–11.74, $p = 0.008$) by rs5417 + rs7658048 + rs12453290 was improved in comparison to the odds ratio of 3.01 (95% CI: 1.44–6.43, $p = 0.004$) by rs5417 alone. However, the SNP–SNP pair (rs5417 + rs3785157) identified by logistic regression did not show improvement over individual SNPs (Table 2).

We further utilized the SNPs identified by MARS and MDR to develop a simple decision tree model for predicting DILI chronicity. As shown in Figure 2, three SNPs (rs6487213, rs5417 and rs12453290) were selected as the predictors for the decision tree model built on 271 patients. The leaf defined by rs6487213, rs5417, with the genotype CC and AA, yielded a frequency of 36.4% (12/33) of DILI chronicity as compared to the overall frequency of 12% (33/271) in the study population. Other leaves defined by rs6487213, rs5417 and rs12453290, together yielded a frequency of 50% (3/6) DILI chronicity. When evaluated by using a stratified 5-fold cross-validation, the decision tree model yields a sensitivity of 42.4%, a specificity of 90.3%, an accuracy of 86.3%, and a balanced accuracy of 66.4%, which is calculated from the average value of sensitivity and specificity [44].

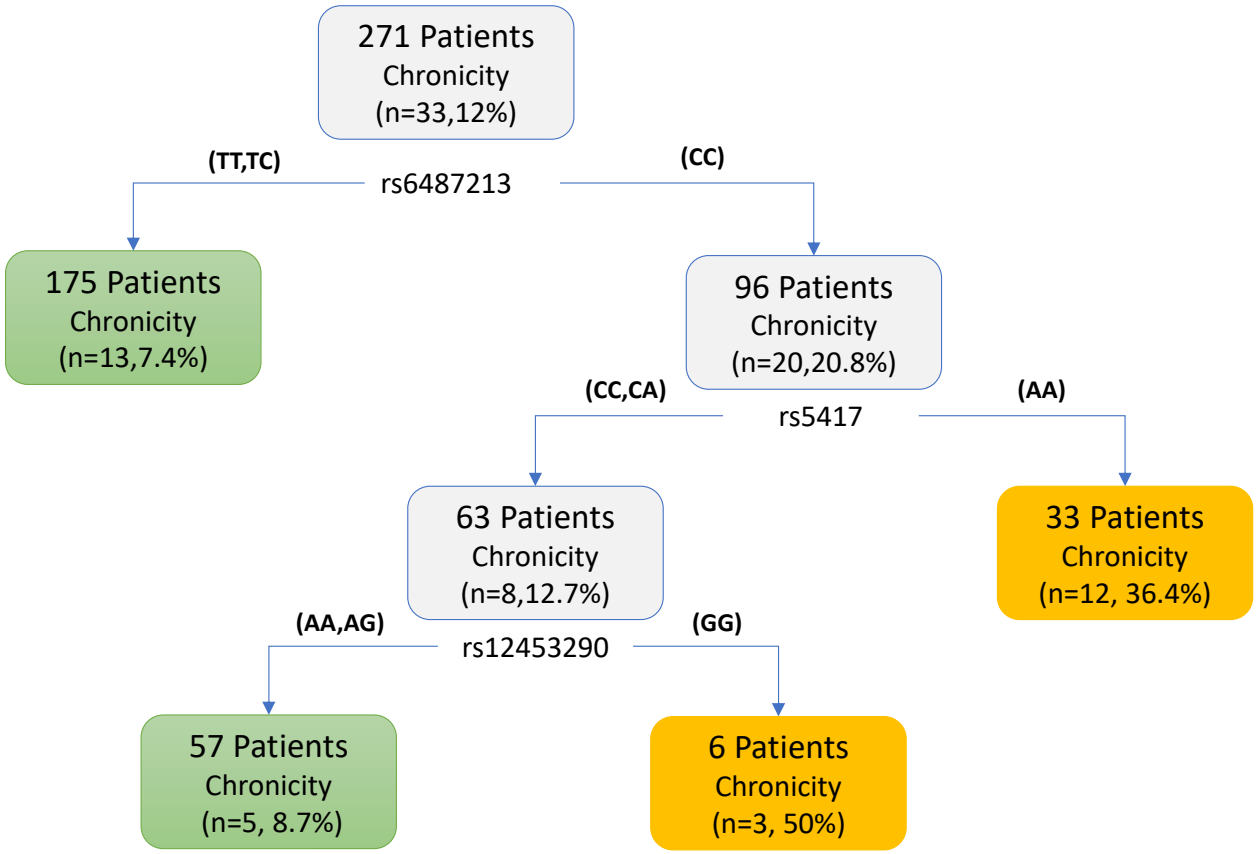

Figure 2. A decision tree model was developed to predict chronicity of drug-induced liver injury for 271 patients. Specifically, in the first layer, the genotype TT and TC of SNP rs6487213 will be assigned as acute, while the genotype CC will be continued to the next layer. In the second layer, the patient with the genotype AA of rs5417 will be assigned as chronic, otherwise, it will need further consideration. In the third layer, if the genotype GG of SNP rs12453290 was determined, the patient will be assigned as chronic and otherwise as acute.

Table 2. Individual SNPs and SNP–SNP interaction analysis for their association with chronic drug-induced liver injury.

Individual SNP Analysis				
Individual SNPs or SNP–SNP Interaction	**Genotypes**	**Acute N (%)**	**Chronicity N (%)**	**Odds Ratio (95% Confidence Intervals, *p* Value)**
rs6487213	CC	76 (79.2)	20 (20.8)	3.28 (1.57–7.09, $p = 0.002$)
	CT or TT	162 (92.6)	13 (7.4)	
rs5417	AA	74 (79.6)	19 (20.4)	3.01 (1.44–6.43, $p = 0.004$)
	CA or CC	164 (92.1)	14 (7.9)	
rs7658048	AG	113 (86.3)	18 (13.7)	1.33 (0.64–2.79, $p = 0.448$)
	AA or GG	125 (89.3)	15 (10.7)	
rs12453290	AA	105 (84.7)	19 (15.3)	1.72 (0.83–3.65, $p = 0.149$)
	GA or GG	133 (90.5)	14 (9.5)	
rs3785157	CC	106 (85.5)	18 (14.5)	1.49 (0.72–3.14, $p = 0.282$)
	TC or TT	132 (89.8)	15 (10.2)	

Table 2. *Cont.*

Individual SNP Analysis				
Individual SNPs or SNP–SNP Interaction	**Genotypes**	**Acute N (%)**	**Chronicity N (%)**	**Odds Ratio (95% Confidence Intervals, *p* Value)**
SNP–SNP interaction analysis				
MARS analysis				
rs6487213 + rs3785157	CC and CC	32 (69.6)	14 (30.4)	4.74 (2.14–10.39, $p < 0.001$)
	Others	206 (91.6)	19 (8.4)	
MDR analysis				
rs5417 + rs7658048 + rs12453290	AA and AG and AA	12 (66.7)	6 (33.3)	4.19 (1.36–11.74, $p = 0.008$)
	Others	226 (89.3)	27 (10.7)	
Random Forest plus logistic regression				
rs5417 + rs3785157	AA and CC	36 (78.3)	10 (21.7)	2.44 (1.03–5.45, $p = 0.034$)
	Others	202 (89.8)	23 (10.2)	

4. Discussion

Individual SNPs reportedly have limited predictive power on human phenotypes, especially for diseases with complicated mechanisms, such as DILI. Here, we investigated three different machine learning approaches: MARS, MDR, and Random Forest plus logistic regression, for their capabilities in identifying SNP–SNP interactions associated with DILI chronicity.

Logistic regression is a widely used approach for detecting SNP–SNP interactions; however, it has limited power to handle the exponentially growing interaction terms or the need of the large sample sizes to detect effects with high dimensions [26]. The combined use of Random Forest and logistic regression improved efficiency in searching SNP interactions but still could miss the best SNP–SNP interaction term by falling into local maxima. MARS and MDR are designed to overcome these limitations to identify the maximum effects of interaction terms in a moderate sample size [37,45]. MDR can detect the high order of SNP interactions, while MARS is faster in revealing candidate SNPs in the chronicity DILI dataset. Note, our DILI chronicity dataset only contains 872 SNPs and is not a large dataset. Regardless, both MARS and MDR successfully identified the SNP–SNP interaction terms with a higher predictive power than individual SNPs, while Random Forest plus logistic regression failed in this test.

A simple predictive model based on the SNPs identified by MARS and MDR also demonstrated a fair predictive performance, suggesting that the presence of these SNP–SNP interactions as risk factors could be associated with an increased risk of chronicity of drug-induced liver injury. We also explored possible underlying biological mechanisms associated with the SNPs identified by the machine learning approaches. Elevated total bilirubin was reportedly associated with DILI chronicity [46,47] and was confirmed in our study. Here, SNP rs6487213, a variant of gene SLCO1B1, was identified as an important contributing factor for predicting chronic DILI. SLCO1B1 is a solute carrier organic anion transporter family member 1B1. It is responsible for bilirubin uptake and transporting xenobiotic compounds such as toxins and drugs from blood into the liver, to be eventually excreted from the body. The uptake capability is markedly decreased by the presence of SLCO1B1 allelic variants. In a study involving 9500 Caucasians, the allelic variation in SLCO1B1 was reported as a major genetic predictor of increased serum bilirubin levels [48]. Furthermore, the polymorphisms of SLCO1B1 were reportedly associated with hepatox-

icity caused by drugs, including but not limited to methimazole [49], methotrexate [50], atorvastatin [51] and anti-tuberculosis drugs [52], including rifampin [53].

Increasing evidence suggests that membrane transporters impact immune cell function and shape immunological responses. The SNP rs5417 is a variant of the gene GLUT4 (also known as solute carrier family 2, facilitated glucose transporter member 4, SLC2A4), which is a critical transporter for glucose uptake and metabolism. GLUT1 upregulation plays a key role in T cell activation and in importing glucose to be metabolized through aerobic glycolysis. Similar to GLUT1, GLUT4 was upregulated in activated CD4+ T cells to varying degrees, and the variants in GLUT4 play compensatory and supporting roles in most T cells [54].

Furthermore, the SNP rs12453290 is a variant of the gene MCT4 (SLC16A3), which is a member of the monocarboxylate transporter (MCT) family controlling the lactate uptake. MCT4 plays a role in exporting additional lactate into the extracellular space. Lactate often builds up at sites of inflammation and in hypoxic conditions. Excess lactate importing into T cells will restrict proliferation and cytokine production of cytotoxic T cells. In T cells, T cell receptor activation upregulates MCT1 expression, while its blockage will result in impaired cytotoxic T cell function. MCT4 upregulation causes full macrophage activation, while its deletion will lead to an intracellular decrease in glycolysis and accumulation of lactate. By contrast, the downregulation of MCT4 will promote the cytotoxicity of natural killer cells and boost immune responses [55].

5. Conclusions

In this study, we investigated three machine learning methods for the study of gene–gene interactions as risk factors for chronic DILI, including MARS, MDR and Random Forest followed by logistic regression. All of the three methods were robust when tested in the simulation dataset with permutation analysis. When we applied these methods to the real-world DILI chronicity dataset, MARS and MDR identified SNP–SNP interactions that showed improved performance regarding the association with DILI chronicity as compared to their individual SNP counterparts. This finding demonstrated that MARS and MDR are superior to the traditional logistic regression approach for identifying SNP–SNP interactions in complex diseases such as chronic DILI.

We also developed a simple decision tree model by using the SNPs from the interactions identified by MARS and MDR, and most of these SNPs were mechanistically relevant, either related to the bilirubin uptake or to immunological responses. These SNPs will be further investigated and validated by including non-genetic factors such as drug properties and responses as co-factors. The identification of SNP–SNP interactions for disease susceptibility has become increasingly important, and our evaluated methods for determining potential genetic risk of the synergetic SNP interactions will potentially benefit the diagnostic test of chronic DILI.

Author Contributions: M.C. conceived and designed the study; R.M., K.A., T.-J.L. and M.C. analyzed the data; M.C. interpreted the data; R.M. and M.C. drafted the manuscript. All authors contributed to the critical review and final approval of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The data was collected by the members in International Serious Adverse Event Consortium, and the study protocols were approved by their local ethics committees.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from International Serious Adverse Event Consortium and are available with the permission of International Serious Adverse Event Consortium.

Acknowledgments: The authors thank the International Serious Adverse Event Consortium (iSAEC) for providing drug-induced liver injury dataset and Joanne Berger, FDA Library, for manuscript editing assistance.

Conflicts of Interest: The authors declare no conflict of interest. This manuscript reflects the views of the author(s) and does not necessarily reflect those of the U.S. Food and Drug Administration.

References

1. Andrade, R.J.; Chalasani, N.; Björnsson, E.S.; Suzuki, A.; Kullak-Ublick, G.A.; Watkins, P.B.; Devarbhavi, H.; Merz, M.; Lucena, M.I.; Kaplowitz, N.; et al. Drug-induced liver injury. *Nat. Rev. Dis. Primers* **2019**, *5*, 1–22. [CrossRef]
2. Hoofnagle, J.H.; Björnsson, E.S. Drug-induced liver injury—Types and phenotypes. *N. Engl. J. Med.* **2019**, *381*, 264–273. [CrossRef]
3. Kaplowitz, N. Idiosyncratic drug hepatotoxicity. *Nat. Rev. Drug Discov.* **2005**, *4*, 489–499. [CrossRef]
4. Chen, M.; Vijay, V.; Shi, Q.; Liu, Z.; Fang, H.; Tong, W. FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov. Today* **2011**, *16*, 697–703. [CrossRef]
5. Chen, M.; Suzuki, A.; Thakkar, S.; Yu, K.; Hu, C.; Tong, W. DILIrank: The largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov. Today* **2016**, *21*, 648–653. [CrossRef] [PubMed]
6. Chalasani, N.; Björnsson, E. Risk factors for idiosyncratic drug-induced liver injury. *Gastroenterology* **2010**, *138*, 2246–2259. [CrossRef] [PubMed]
7. European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu; Clinical Practice Guideline Panel: Chair; Panel members; EASL Governing Board representative. EASL clinical practice guidelines: Drug-induced liver injury. *J. Hepatol.* **2019**, *70*, 1222–1261. [CrossRef] [PubMed]
8. Kaplowitz, N. Drug-induced liver injury. *Clin. Infect. Dis.* **2004**, *38* (Suppl. 2), S44–S48. [CrossRef] [PubMed]
9. Chen, M.; Suzuki, A.; Borlak, J.; Andrade, R.J.; Lucena, M.I. Drug-induced liver injury: Interactions between drug properties and host factors. *J. Hepatol.* **2015**, *63*, 503–514. [CrossRef] [PubMed]
10. Amacher, D.E. The primary role of hepatic metabolism in idiosyncratic drug-induced liver injury. *Expert Opin. Drug Metab. Toxicol.* **2012**, *8*, 335–347. [CrossRef]
11. Stephens, C.; Andrade, R.J. Genetic predisposition to drug-induced liver injury. *Clin. Liver Dis.* **2020**, *24*, 11–23. [CrossRef]
12. Hoofnagle, J.H.; Bonkovsky, H.L.; Phillips, E.J.; Li, Y.J.; Ahmad, J.; Barnhart, H.; Durazo, F.; Fontana, R.J.; Gu, J.; Khan, I.; et al. HLA-B*35:01 and Green Tea–Induced Liver Injury. *Hepatology* **2021**, *73*, 2484–2493. [CrossRef]
13. Kaliyaperumal, K.; Grove, J.I.; Delahay, R.M.; Griffiths, W.J.H.; Duckworth, A.; Aithal, G.P. Pharmacogenomics of drug-induced liver injury (DILI): Molecular biology to clinical applications. *J. Hepatol.* **2018**, *69*, 948–957. [CrossRef]
14. Li, Y.J.; Phillips, E.J.; Dellinger, A.; Nicoletti, P.; Schutte, R.; Li, D.; Ostrov, D.A.; Fontana, R.J.; Watkins, P.B.; Stolz, A.; et al. Human leukocyte antigen B*14:01 and B*35:01 are associated with trimethoprim-sulfamethoxazole induced liver injury. *Hepatology* **2021**, *73*, 268–281. [CrossRef]
15. Fontana, R.J.; Cirulli, E.T.; Gu, J.; Kleiner, D.; Ostrov, D.; Phillips, E.; Schutte, R.; Barnhart, H.; Chalasani, N.; Watkins, P.B.; et al. The role of HLA-A*33:01 in patients with cholestatic hepatitis attributed to terbinafine. *J. Hepatol.* **2018**, *69*, 1317–1325. [CrossRef] [PubMed]
16. Urban, T.J.; Nicoletti, P.; Chalasani, N.; Serrano, J.; Stolz, A.; Daly, A.K.; Aithal, G.P.; Dillon, J.; Navarro, V.; Odin, J.; et al. Minocycline hepatotoxicity: Clinical characterization and identification of HLA-B∗35:02 as a risk factor. *J. Hepatol.* **2017**, *67*, 137–144. [CrossRef] [PubMed]
17. Daly, A.K.; Donaldson, P.T.; Bhatnagar, P.; Shen, Y.; Pe'er, I.; Floratos, A.; Daly, M.J.; Goldstein, D.B.; John, S.; Nelson, M.R.; et al. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat. Genet.* **2009**, *41*, 816–819. [CrossRef] [PubMed]
18. Nicoletti, P.; Aithal, G.P.; Bjornsson, E.S.; Andrade, R.J.; Sawle, A.; Arrese, M.; Barnhart, H.X.; Bondon-Guitton, E.; Hayashi, P.H.; Bessone, F.; et al. Association of liver injury from specific drugs, or groups of drugs, with polymorphisms in HLA and other genes in a genome-wide association study. *Gastroenterology* **2017**, *152*, 1078–1089. [CrossRef]
19. Cirulli, E.T.; Nicoletti, P.; Abramson, K.; Andrade, R.J.; Bjornsson, E.S.; Chalasani, N.; Fontana, R.J.; Hallberg, P.; Li, Y.J.; Lucena, M.I.; et al. A missense variant in PTPN22 is a risk factor for drug-induced liver injury. *Gastroenterology* **2019**, *156*, 1707–1716.e2. [CrossRef]
20. Urban, T.J.; Shen, Y.; Stolz, A.; Chalasani, N.; Fontana, R.J.; Rochon, J.; Ge, D.; Shianna, K.V.; Daly, A.K.; Lucena, M.I.; et al. Limited contribution of common genetic variants to risk for liver injury due to a variety of drugs. *Pharm. Genom.* **2012**, *22*, 784. [CrossRef]
21. Overby, C.L.; Hripcsak, G.; Shen, Y. Estimating heritability of drug-induced liver injury from common variants and implications for future study designs. *Sci. Rep.* **2014**, *4*, 1–3. [CrossRef] [PubMed]
22. Carlborg, Ö.; Haley, C.S. Epistasis: Too often neglected in complex trait studies? *Nat. Rev. Genet.* **2004**, *5*, 618–625. [CrossRef] [PubMed]
23. Yi, N. Statistical analysis of genetic interactions. *Genet. Res.* **2010**, *92*, 443–459. [CrossRef]
24. Wei, W.-H.; Hemani, G.; Haley, C.S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **2014**, *15*, 722–733. [CrossRef]
25. Cook, N.R.; Zee, R.Y.; Ridker, P.M. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat. Med.* **2004**, *23*, 1439–1453. [CrossRef]

26. Moore, J.H.; Williams, S.M. New strategies for identifying gene-gene interactions in hypertension. *Ann. Med.* **2002**, *34*, 88–95. [CrossRef]
27. Bansal, A.; Pepe, M.S. When does combining markers improve classification performance and what are implications for practice? *Stat. Med.* **2013**, *32*, 1877–1892. [CrossRef]
28. Aithal, G.P.; Watkins, P.B.; Andrade, R.J.; Larrey, D.; Molokhia, M.; Takikawa, H.; Hunt, C.M.; Wilke, R.A.; Avigan, M.; Kaplowitz, N.; et al. Case definition and phenotype standardization in drug-induced liver injury. *Clin. Pharmacol. Ther.* **2011**, *89*, 806–815. [CrossRef]
29. Lucena, M.I.; Molokhia, M.; Shen, Y.; Urban, T.J.; Aithal, G.P.; Andrade, R.J.; Day, C.P.; Ruiz-Cabello, F.; Donaldson, P.T.; Stephens, C.; et al. Susceptibility to amoxicillin-clavulanate-induced liver injury is influenced by multiple HLA class I and II alleles. *Gastroenterology* **2011**, *141*, 338–347. [CrossRef]
30. Marees, A.T.; de Kluiver, H.; Stringer, S.; Vorspan, F.; Curis, E.; Marie-Claire, C.; Derks, E.M. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **2018**, *27*, e1608. [CrossRef] [PubMed]
31. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdóttir, H.; Tamayo, P.; Mesirov, J.P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011**, *27*, 1739–1740. [CrossRef] [PubMed]
32. Barron, A.R.; Xiao, X. Discussion: Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 67–82. [CrossRef]
33. Ritchie, M.D.; Hahn, L.W.; Roodi, N.; Bailey, L.R.; Dupont, W.D.; Parl, F.F.; Moore, J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **2001**, *69*, 138–147. [CrossRef] [PubMed]
34. Ritchie, M.D.; Motsinger, A.A. Multifactor dimensionality reduction for detecting gene–gene and gene–environment interactions in pharmacogenomics studies. *Pharmacogenomics* **2005**, *6*, 823–834. [CrossRef]
35. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
36. Cho, Y.M.; Ritchie, M.D.; Moore, J.H.; Park, J.Y.; Lee, K.U.; Shin, H.D.; Lee, H.K.; Park, K.S. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* **2004**, *47*, 549–554. [CrossRef]
37. Lin, H.Y.; Wang, W.; Liu, Y.H.; Soong, S.J.; York, T.P.; Myers, L.; Hu, J.J. Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP–SNP interactions and their application in prostate cancer. *J. Hum. Genet.* **2008**, *53*, 802–811. [CrossRef]
38. Goldstein, B.A.; Polley, E.C.; Briggs, F.B. Random forests for genetic association studies. *Stat. Appl. Genet. Mol. Biol.* **2011**, *10*, 32. [CrossRef]
39. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013; Available online: http://www.R-project.org/ (accessed on 5 October 2021).
40. Winham, S.J.; Motsinger-Reif, A.A. An R package implementation of multifactor dimensionality reduction. *BioData Min.* **2011**, *4*, 24. [CrossRef] [PubMed]
41. Breiman, L.; Cutler, A. Breiman and Cutler's Random Forests for Classification and Regression. Package 'randomForest' Published on CRAN. 2018. Available online: https://www.stat.berkeley.edu/~{}breiman/RandomForests/ (accessed on 5 October 2021).
42. Therneau, T.; Atkinson, B.; Ripley, B. Rpart: Recursive Partitioning. R Package Version 4.1-3. 2019. Available online: http://CRAN.R-project.org/package=rpart (accessed on 5 October 2021).
43. Minitab 17 Statistical Software. Minitab, Inc.: State College, PA, USA, 2010. Available online: https://www.minitab.com (accessed on 5 October 2021).
44. García, V.; Mollineda, R.A.; Sánchez, J.S. Index of balanced accuracy: A performance measure for skewed class distributions. In Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2009), Póvoa de Varzim, Portugal, 10–12 June 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 441–448.
45. Gui, J.; Moore, J.H.; Williams, S.M.; Andrews, P.; Hillege, H.L.; van der Harst, P.; Navis, G.; Van Gilst, W.H.; Asselbergs, F.W.; Gilbert-Diamond, D. A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PLoS ONE* **2013**, *8*, e66545. [CrossRef]
46. Fontana, R.J.; Hayashi, P.H.; Barnhart, H.; Kleiner, D.E.; Reddy, K.R.; Chalasani, N.; Lee, W.M.; Stolz, A.; Phillips, T.; Serrano, J.; et al. Persistent liver biochemistry abnormalities are more common in older patients and those with cholestatic drug induced liver injury. *Am. J. Gastroenterol.* **2015**, *110*, 1450. [CrossRef] [PubMed]
47. Medina-Caliz, I.; Robles-Diaz, M.; Garcia-Muñoz, B.; Stephens, C.; Ortega-Alonso, A.; Garcia-Cortes, M.; González-Jimenez, A.; Sanabria-Cabrera, J.A.; Moreno, I.; Fernandez, M.C.; et al. Definition and risk factors for chronicity following acute idiosyncratic drug-induced liver injury. *J. Hepatol.* **2016**, *65*, 532–542. [CrossRef] [PubMed]
48. Johnson, A.D.; Kavousi, M.; Smith, A.V.; Chen, M.H.; Dehghan, A.; Aspelund, T.; Lin, J.P.; van Duijn, C.M.; Harris, T.B.; Cupples, L.A.; et al. Genome-wide association meta-analysis for total serum bilirubin levels. *Hum. Mol. Genet.* **2009**, *18*, 2700–2710. [CrossRef] [PubMed]
49. Yang, L.; Yang, L.; Wu, H.; Gelder, T.V.; Matic, M.; Ruan, J.S.; Han, Y.; Xie, R.X. SLCO1B1 rs4149056 genetic polymorphism predicting methotrexate toxicity in Chinese patients with non-Hodgkin lymphoma. *Pharmacogenomics* **2017**, *18*, 1557–1562. [CrossRef] [PubMed]
50. Jin, S.; Li, X.; Fan, Y.; Fan, X.; Dai, Y.; Lin, H.; Cai, W.; Yang, J.; Xiang, X. Association between genetic polymorphisms of SLCO1B1 and susceptibility to methimazole-induced liver injury. *Basic Clin. Pharmacol. Toxicol.* **2019**, *125*, 508–517. [CrossRef]

51. Shu, N.; Hu, M.; Ling, Z.; Liu, P.; Wang, F.; Xu, P.; Zhong, Z.; Sun, B.; Zhang, M.; Li, F.; et al. The enhanced atorvastatin hepatotoxicity in diabetic rats was partly attributed to the upregulated hepatic Cyp3a and SLCO1B1. *Sci. Rep.* **2016**, *6*, 33072. [CrossRef]
52. Chen, R.; Wang, J.; Tang, S.; Zhang, Y.; Lv, X.; Wu, S.; Xia, Y.; Deng, P.; Ma, Y.; Tu, D.; et al. Association of polymorphisms in drug transporter genes (SLCO1B1 and SLC10A1) and anti-tuberculosis drug-induced hepatotoxicity in a Chinese cohort. *Tuberculosis* **2015**, *95*, 68–74. [CrossRef]
53. Li, L.M.; Chen, L.; Deng, G.H.; Tan, W.T.; Dan, Y.J.; Wang, R.Q.; Chen, W.S. SLCO1B1 *15 haplotype is associated with rifampin-induced liver injury. *Mol. Med. Rep.* **2012**, *6*, 75–82.
54. Weiss, H.J.; Angiari, S. Metabolite Transporters as Regulators of Immunity. *Metabolites* **2020**, *10*, 418. [CrossRef]
55. Long, Y.; Gao, Z.; Hu, X.; Xiang, F.; Wu, Z.; Zhang, J.; Han, X.; Yin, L.; Qin, J.; Lan, L.; et al. Downregulation of MCT4 for lactate exchange promotes the cytotoxicity of NK cells in breast carcinoma. *Cancer Med.* **2018**, *7*, 4690–4700. [CrossRef]