American Journal of Transplantation 2016; XX: 1–12
Wiley Periodicals Inc.

doi: 10.1111/ajt.14040

Predictive Modeling of Tacrolimus Dose Requirement Based on High-Throughput Genetic Screening

C. Damon[1,*], M. Luck[1,2], L. Toullec[3], I. Etienne[4], M. Buchler[5], B. Hurault de Ligny[6], G. Choukroun[7], A. Thierry[8], C. Vigneau[9], B. Moulin[10], A.-E. Heng[11], J.-F. Subra[12], C. Legendre[13], A. Monnot[1], A. Yartseva[1], M. Bateson[1], P. Laurent-Puig[2,3,14], D. Anglicheau[13], P. Beaune[2,3,14], M. A. Loriot[2,3,14], E. Thervet[2,15] and N. Pallet[2,3,14,15,*]

[1]*Hypercube Institute, Paris, France*
[2]*Paris Descartes University, Paris, France*
[3]*Department of Clinical Chemistry, Georges Pompidou European Hospital, Assistance Publique Hôpitaux de Paris, Paris, France*
[4]*Department of Nephrology, Rouen University Hospital, Rouen, France*
[5]*Department of Nephrology, Tours University Hospital, Tours, France*
[6]*Department of Nephrology, Caen University Hospital, Caen, France*
[7]*Department of Nephrology, Amiens University Hospital, Amiens, France*
[8]*Department of Nephrology, Poitiers University Hospital, Poitiers, France*
[9]*Department of Nephrology, Rennes University Hospital, Rennes, France*
[10]*Department of Nephrology, Strasbourg University Hospital, Strasbourg, France*
[11]*Department of Nephrology, Clermont-Ferrand University Hospital, Clermont-Ferrand, France*
[12]*Department of Nephrology, Angers University Hospital, Angers, France*
[13]*Department of Nephrology, Necker Hospital, Assistance Publique Hôpitaux de Paris, Paris, France*
[14]*Institut National pour la Santé et la Recherche Médicale (INSERM) U1147, Paris, France*
[15]*Department of Nephrology, Georges Pompidou European Hospital, Assistance Publique Hôpitaux de Paris, Paris, France*
**Corresponding authors: Nicolas Pallet and Cécilia Damon, nicolas.pallet@aphp.fr and ceciliadamon@gmail.com*

Any biochemical reaction underlying drug metabolism depends on individual gene–drug interactions and on groups of genes interacting together. Based on a high-throughput genetic approach, we sought to identify a set of covariant single-nucleotide polymorphisms predictive of interindividual tacrolimus (Tac) dose requirement variability. Tac blood concentrations (Tac C_0) of 229 kidney transplant recipients were repeatedly monitored after transplantation over 3 mo. Given the high dimension of the genomic data in comparison to the low number of observations and the high multicolinearity among the variables (gene variants), we developed an original predictive approach that integrates an ensemble variable-selection strategy to reinforce the stability of the variable-selection process and multivariate modeling. Our predictive models explained up to 70% of total variability in Tac C_0 per dose with a maximum of 44 gene variants (p-value <0.001 with a permutation test). These models included molecular networks of drug metabolism with oxidoreductase activities and the multidrug-resistant ABCC8 transporter, which was found in the most stringent model. Finally, we identified an intronic variant of the gene encoding SLC28A3, a drug transporter, as a key gene involved in Tac metabolism, and we confirmed it in an independent validation cohort.

Abbreviations: ABCC8, ATP binding cassette C8; ANOVA, Analysis of variance; CV, cross-validation; CYB5R3, cytochrome b5 reductase 3; CYP, cytochrome P450; KTR, kidney transplant recipient; log (Tac C_0/dose), logarithmically transformed tacrolimus blood concentration per dose; MI, mutual information; PLS, partial least squares; SE, standard error; SLC28A3, solute carrier family 28A3; SNP, single-nucleotide polymorphism; Tac, tacrolimus; Tac C_0, tacrolimus blood concentration; Tac C_0/dose, tacrolimus blood concentration per dose

Received 29 June 2016, revised 24 August 2016 and accepted for publication 26 August 2016

Introduction

Genetic factors have been estimated to account for a major part of interindividual differences in drug metabolism and response and to have the most important influence on drug-treatment outcomes for some drugs (1,2). With the availability of the human genome sequence and technologies that allow high-throughput genotyping, the pharmacogenomic approach has provided insights into patient selection for specific therapies and dosages (3). This approach is particularly relevant for drugs with a narrow therapeutic index and great interindividual variability, such as immunosuppressive agents, because it may

increase the likelihood of successful treatment while reducing the risk of adverse effects (4).

The most critical genetic polymorphisms associated with variations in the tacrolimus (Tac) pharmacokinetic profile affect the genes encoding cytochrome P450 (CYP) 3A4 and CYP3A5 enzymes, which are involved in intestinal and hepatic metabolism of Tac (5–8). *CYP3A4* and *CYP3A5* variants, however, explain only part of the variation in Tac bioavailability, suggesting the involvement of a wider network of candidate genes. Other candidates likely explain the genetic basis for interindividual variability in dose-adjusted Tac blood concentrations (Tac C_0). Given the high level of interdependence between individual genes, we can assume that any biochemical reactions underlying drug metabolism could not depend on gene–drug interactions at the individual level but rather on a group of genes interacting with each other. No exhaustive exploration of the metabolic networks controlling Tac metabolism and how their alterations affect pharmacokinetic parameters has been performed to date.

To address this issue, we performed high-throughput screening with the aim of identifying a set of covariant single-nucleotide polymorphisms (SNPs) predictive of the interpatient Tac dose-requirement variability in a population of kidney transplant recipients (KTRs). Because of the high dimension of the genomic data in comparison to the relatively low number of observations and the high multicolinearity among the variables (SNPs), we developed an original predictive approach that integrates an ensemble variable-selection strategy—based on a Fisher exact test and a measure of mutual Information (MI) to reinforce the stability of the feature selection process—and multivariate modeling (the partial least squares [PLS] multivariate predictive method) (9–11).

Patients and Methods

Patients

Study cohort: The high-throughput DNA genotyping analysis was performed on the Tactique study cohort (12). Initially, the Tactique study was conducted to evaluate whether adaptation of Tac dosing according to the *CYP3A5* genotype would allow earlier achievement of target Tac C_0 in KTRs. The design of the study was detailed by Thervet et al (12). Overall, 280 KTRs from 12 sites in France were randomly assigned at day 7 after transplantation to receive Tac (twice-daily formulation; Prograf; Astellas, Tokyo, Japan) at a dosage that was fixed at 0.2 mg/kg per day (control group) or that was determined by individual genotype. Patients who expressed CYP3A5 (carriers of at least one *CYP3A5*1* allele) received 0.3 mg/kg per day, whereas patients who did not express CYP3A5 (*CYP3A5*3/*3* genotype) received 0.15 mg/kg per day (adapted-dose group). All patients received a biological induction with basiliximab (19%) or rabbit thymocyte antiglobulin (81%) and received 3 g mycophenolate mofetil (Cell-Cept; Roche, Basel, Switzerland) daily for 15 days (tapered to 2 g/day) and a tapered corticosteroid regimen as the maintenance regimen. The first measurement of Tac C_0 was performed after the intake of six doses (corresponding to day 10 after transplantation), after which physicians could modify the daily dose to achieve a prespecified Tac C_0 target range between 10 and 15 ng/mL. At this time, the proportion of patients within the therapeutic target Tac C_0 between the two groups (control and adapted-dose group) was calculated to determine the primary end point. After that, Tac C_0 was recorded at days 14, 30, 60, and 90 after transplantation. Among the 280 patients of the initial study, DNA samples were available for 272 patients; among them, genotyping was successfully done on the SNP microarray for 229 patients without any missing clinical or biological values over the entire follow-up period. Our descriptive study is essentially based on the whole cohort (and thus is independent of the randomization arms), and we did not take into account the primary end point of the Tactique study (percentage of patients reaching the prespecified Tac C_0 range) but rather assessed Tac C_0 per dose (Tac C_0/dose) over time as the explanatory variable; therefore, a nonrandom loss of sample would likely have no impact on our conclusions. The demographic and clinical characteristics of this cohort of 229 patients are described in Table S1. Tac C_0 was measured using the EMIT 2000 Tac assay (Siemens, Munich, Germany) (13).

Validation cohort: To confirm the association between the *SLC28A3* gene variant (rs10868152) and Tac dose response, we analyzed an independent cohort of 189 KTRs from Necker Hospital for whom DNA was available for genotyping and who had provided informed consent. The demographic and clinical information was collected prospectively in the DIVAT (*Données Informatisées et VAlidées en Transplantation*) database at 3, 12 and 24 mo after transplantation. The demographic and clinical characteristics of this cohort are described in Table S2.

Regulatory aspects: The institutional review board at each participating center approved the study design, and written informed consent was obtained from all patients. The clinical and research activities reported were consistent with the principles of the Declaration of Istanbul, as outlined in the Declaration of Istanbul on Organ Trafficking and Transplant Tourism.

Genotyping and annotation

Genomic DNA was purified using the QIAamp DNA purification system (Qiagen, Hilden, Germany). DNA samples were genotyped for 16 561 SNPs using a customized Illumina SNP genotyping assay (Illumina, San Diego, CA) designed to capture the genetic variation of 1653 key drug pathway genes (including phase I and II drug metabolism enzymes, drug transporters, drug targets, and drug receptors). SNP selection was performed by tagging functional SNPs with tagSNP using the HapMap database (14) with a pairwise tagging cutoff of R^2=0.8. SNPs were selected to characterize the main haplotypes within the white population (95% of haplotypic diversity) according to the following criteria: Genes were defined by their position on human genome 36 (National Center for Biotechnology Information); the minor allelic frequency had to be 5%; and for SNPs carrying the same information, a "score design" allowed selection of the final SNPs (defined by technical criteria related to the chip and established by the manufacturer). Microarrays were processed by Integragen (Evry, France) using the Illumina technology and Infinium iSelect custom genotyping (Illumina).

Data mining methodology

Because of the high dimension of the genomic data in comparison to the low number of observations and the high multicolinearity among the variables (SNPs), we developed an original predictive approach associating an ensemble variable-selection scheme and an explanatory soft model. Ensemble methods associate multiple learning algorithms to achieve better predictive performance that could be obtained from any of the

constituent learning algorithms. We combined two complementarity strategies of univariate filtering techniques: a Fisher exact test based on a univariate linear model and a measure of MI. For the explanatory multivariate model, we applied the PLS multivariate predictive method because it allows extraction of groups of genes from among high-dimensional and correlated data that will jointly explain the variation of Tac drug response at each follow-up time after transplantation and over all recorded times. From high-dimensional correlated data, our algorithm aims to extract the relevant groups of SNPs underlying the biochemical processes responsible for the variation in drug responses as the logarithmically transformed Tac C_0/dose (log [Tac C_0/dose]). Tac C_o/dose is a relevant index of dose requirement for drugs requiring dose adjustment based on predose blood levels; because its distribution was skewed, we log-transformed it to obtain a Gaussian distribution.

Ensemble variable-selection strategy: Our ensemble variable-selection strategy combined two complementarity strategies of univariate filtering techniques: a Fisher exact test and a measure of MI. The Fisher exact test based on a univariate linear model allows testing of the linear effect of a single regressor of interest (target: log [Tac C_0/dose] sequentially for many explanatory regressors [features: SNPs]) (Data S1). MI is complementary to the Fisher exact test because it measures the strength of the dependencies (targets and feature) with nonparametric implementation and makes no prior assumptions about the distribution of the data. Moreover, MI accounts for high-order statistics and bounds the optimal Bayes error rate. To avoid the discretization of the data and the high computational cost of numerical integration necessary to assess the probability density functions of the continuous variables, we used an alternative MI estimator. This estimator is based on Renyi quadratic entropy and Cauchy–Schwartz divergence combined with the Parzen window density estimator for continuous variables (15–17) (Data S1).

The combined subset of variables arises from the intersection of the two sets of variables selected by the Fisher exact test and MI methods. For each method, we tested a range of 10 evenly spaced thresholds from 0.001 to 0.1 for normalized MI scores and p-values of 0.05, 0.04, 0.03, 0.02, 0.01, 0.009, 0.008, 0.007, 0.006, 0.005, 0.004, 0.003, 0.002, and 0.001 for the Fisher exact test. When we considered all times together, the selection was based on the average of the Fisher exact test p-values and normalized MI scores computed over all follow-up times after transplantation. The best couple of thresholds (Fisher exact test p-value, normalized MI score) was optimized with inner 10-fold cross-validation such that the best couple was the smallest subset of variables corresponding to a model having an error rate (multivariate coefficient of determination [R^2 value], described in "Evaluation") of no more than 1 standard error (SE) worse than the best model according to the 1-SE rule (18).

Temporal explanatory soft modeling: For the regression model, we used the PLS regression multivariate predictive method because it deals with multidimensional and highly correlated data and requires relatively few observations (19). This soft modeling approach used an iterative greedy procedure to extract the latent component pair from the multidimensional input (i.e. SNPs) and output data (log [Tac C_0/dose], the target variable) that is maximally correlated. Consequently, it handles both overfitting and the identification of explanatory gene groups (i.e. linear combinations of SNPs) responsible for or predictive of the variation of drug response at a given time and over all recorded times through minimization of the least squares error. In our study, we empirically fixed the number of components of our PLS models at two: PLS1 had a one-dimensional response (i.e. models at each follow-up time after transplantation), and PLS2 had a multidimensional response (i.e. model over the recorded times).

Evaluation: We split the data set into training (80%) and test (20%) sets in which the distribution of the target variable was preserved. The training set of 80% of the whole study group was used to learn models, and prediction (i.e. evaluation of model performance) was assessed with the test set (i.e. the remaining 20% of the study group not used for learning) (Figure 1). Of note, the final subset of selected variables was optimized with an inner 10-fold cross-validation and selection criteria based on the R^2 value and the 1-SE rule. The predictive performance of the models was assessed with the R^2 value. The R^2 value represents the proportion of explained variance of the target in which $R^2 = 1$ indicates perfect prediction of the data by the model. For each PLS regression model, a permutation test of 1000 runs was performed to assess the statistical significance of the predictive performance of the model (i.e. $p \leq 0.05$) by comparing its performance with 1000 other PLS regression models built from the same initial data set split. Interestingly, the PLS model assigns a weight to each variable determining its contribution to target prediction.

Having obtained the different models and their sets of explanatory variables, we focused on the statistical effect of each SNP present in the models on the Tac metabolism variations, considering the three possible polymorphisms over time. To do this, we used both repeated-measures analysis of variance (ANOVA) and a linear mixed model to consider fixed and random effects. We built our linear mixed models with the genotype, time, and genotype-by-time variables as fixed effects and the patient as a random effect to evaluate effects of genotype, of time, and of both factors on the mean Tac response (log [Tac C_0/dose]) over time.

Results

Multivariate models predictive of the interindividual variability of Tac dose requirement

We generated models of gene interaction that were predictive of the dose requirement for Tac, using log (Tac C_0/dose) as an index of dose requirement based on predose blood levels. We built two types of models: PLS1, which predicts the variability of the dose requirement at each period of time after transplantation, and PLS2, which predicts variability over the whole follow-up period. The performance, significance and complexity of the models are summarized in Table 1. All performances were significant, with a p-value <0.001 assessed with a permutation test.

Considering the PLS1 models (Table 1 and Figure 2), the best performances were reached at days 60 and 90, with 70.2% and 62.9%, respectively, of the explained variance. These models were also the most complex, with 44 and 33 genes, respectively. This is likely because Tac C_0 and doses are more stable at later time points after transplantation, generating less variance and thus less background noise and allowing the identification of more complete gene interaction networks. In addition, the number of factors that affect Tac bioavailability tends to fall progressively as time after transplantation increases. Consequently, global variability of dose response is expected to increase as part of gene interaction. As expected, *CYP3A5* contributed to the models at all time periods after transplantation, and *CYP3A4* contributed to

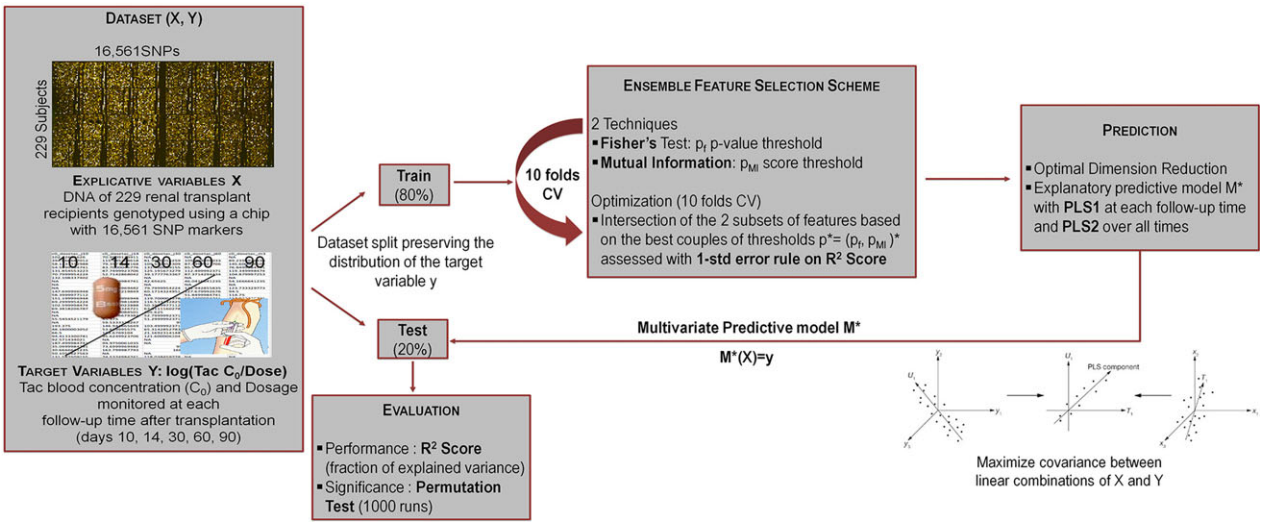

Figure 1: Data-mining methodology. Our predictive approach has two steps. In step 1, an ensemble feature-selection strategy combines two complementary univariate filtering techniques: a Fisher exact test and a measure of mutual information. Variable selection is optimized with an inner 10-fold cross-validation and selection criteria based on the R^2 value and the 1-standard error rule. In step 2, an explanatory soft model uses PLS regression: PLS1 indicates the model for each time and PLS2 indicates the model over all times. The data set is split into a training set (80%) and a test set (20%). Model evaluation relies on the R^2 value and a permutation test of 1000 runs. CV, cross-validation; log (Tac C_0/dose), logarithmically transformed tacrolimus blood concentration per dose; PLS, partial least squares; SNP, single-nucleotide polymorphism.

Table 1: Performance, statistical significance, and complexity of the predictive models at each follow-up time after transplantation (days 10, 14, 30, 60, 90) with PLS1 model and for all times with PLS2 model

	PLS1 models					PLS2 models
Time after transplantation (days)	10	14	30	60	90	10, 14, 30, 60, and 90 days
Performance (R^2 value)	0.30	0.27	0.41	0.7	0.62	0.28
Significance (p-value)	0.001	0.001	0.001	0.001	0.001	0.001
Model complexity (number of SNPs)	5	19	12	44	33	7

PLS, partial least squares; SNP, single-nucleotide polymorphism.

the models at all time periods but one, which confirmed previous findings on the impact of gene variation in the *CYP3A* family in response to Tac and validated our predictive modeling algorithm (Figure 2). Consistent with this result, molecular interaction network analysis using the genes involved in PLS1 models indicated that the most enriched pathways in molecular interaction networks are related to oxidoreductase functions and monooxygenase activity, suggesting that the response to Tac is mediated mostly by drug-metabolizing enzymes (Figure 3).

The PLS2 model, which is predictive of log (Tac C_0/dose) throughout the entire follow-up period after transplantation (Table 1), included only seven SNPs corresponding to *CYP3A4, CYP3A5, CYP3A7* (a pseudogene), *CYP3A43* and ATP binding cassette C8 (*ABCC8*), a multidrug transporter, and explained 28.2% of the variance. This simple model included few genes that were likely to have a more robust impact on Tac metabolism because the selected genes predicted Tac response throughout the posttransplantation period (different from genes present in PLS1 models, which are independent from each other at specific periods of time). Interestingly, variants of *ABCC8* have been associated with Tac-induced new-onset diabetes after transplantation (20). Of note, the models generated for the all-white population yielded results similar to the whole population, with slightly higher complexity and prediction performance (data not shown), indicating that the inclusion of nonwhite participants in the cohort did not affect the prediction of the models.

Notably, in a subgroup analysis, the PLS2 model as applied only to the control group (patients with a fixed Tac dose, described in "Patients and Methods") explained >60% of the target response over all follow-up periods based on 158 selected SNPs (Table S3). In

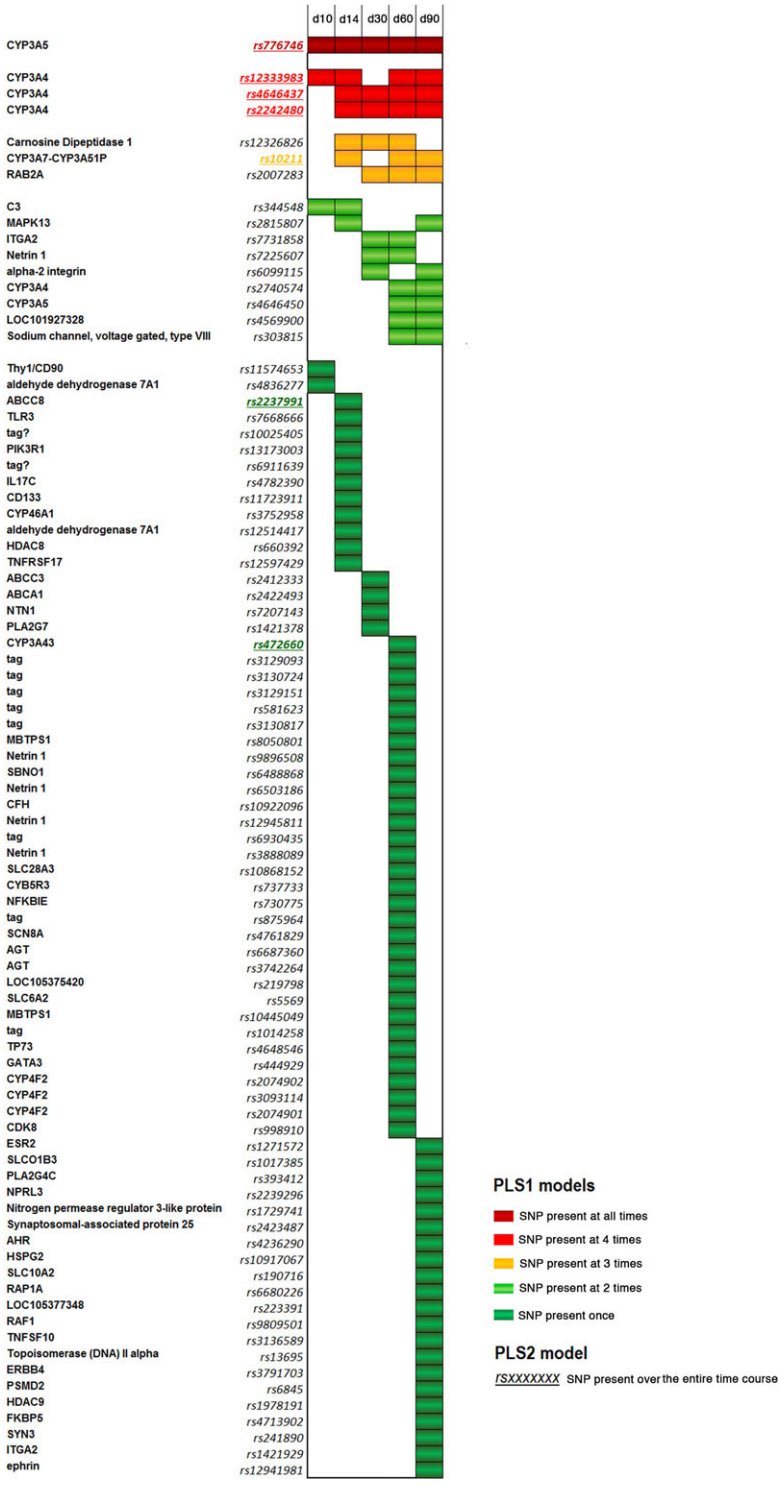

Figure 2: Genes involved in the predictive models of log (Tac C_0/dose) at each follow-up time after transplantation (PLS1 model) and throughout the follow-up period (PLS2 model). Colors represent the number of PLS1 models to which a particular SNP belongs: The SNP is present in five (dark red), four (orange), three (yellow), two (light green), or one model (dark green). SNPs contributing to the PLS2 model (i.e. predicts drug response over the entire time course after transplantation) are underlined. d, day; log (Tac C_0/dose), logarithmically transformed tacrolimus blood concentration per dose; PLS, partial least squares; SNP, single-nucleotide polymorphism.

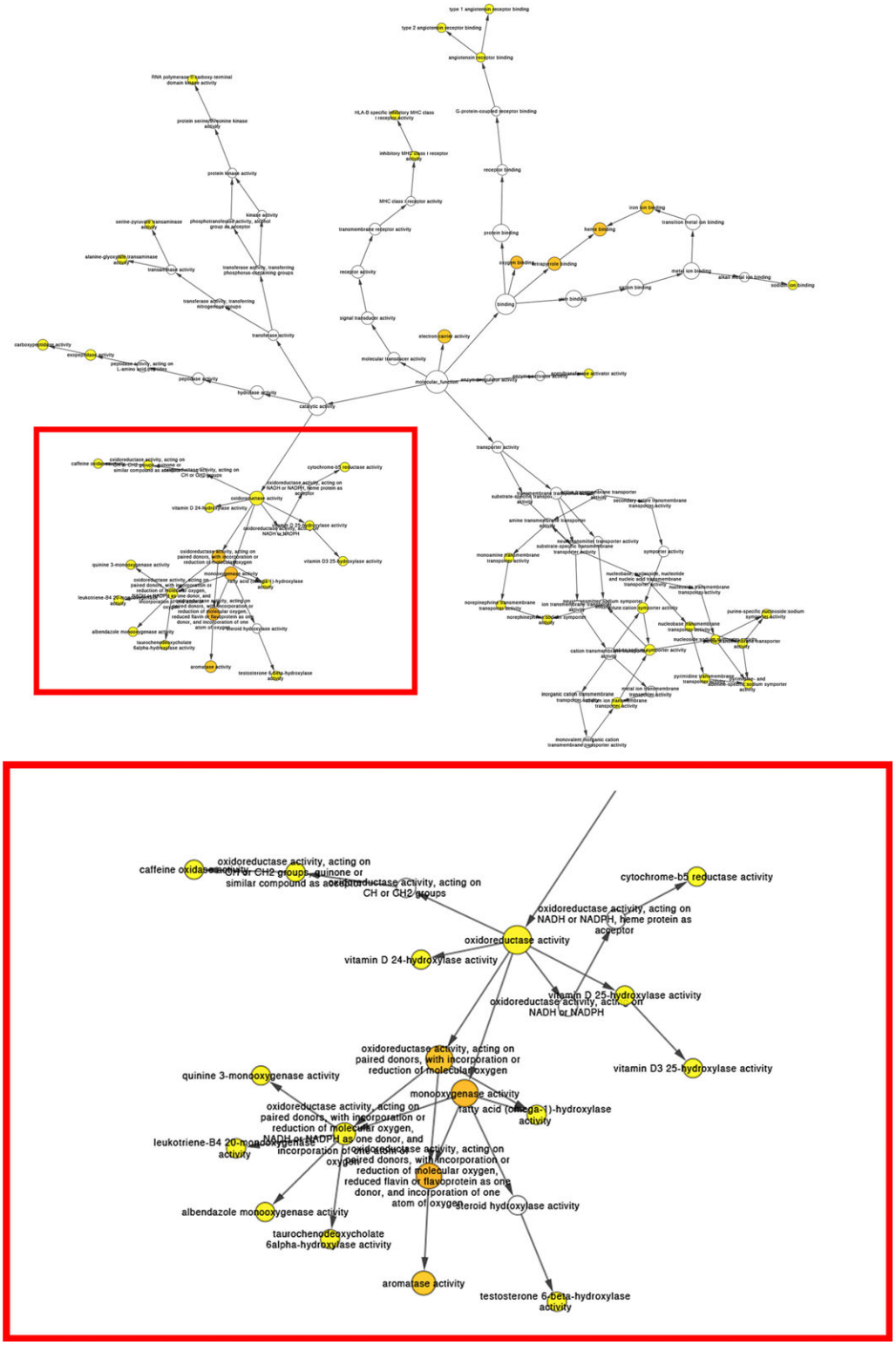

Figure 3: Biological processes enriched by genes identified by predictive models of logarithmically transformed tacrolimus blood concentration per dose at each follow-up time after transplantation. Gene ontology (biological function pathway enrichment) performed using Cytoscape for all genes selected by the predictive models. Genes implicated in these models are related to drug metabolism and are in molecular interaction networks related to monooxygenase functions.

contrast, the model explained only ≈7% of the response based on 32 selected SNPs when applied to the *CYP3A5*3/*3* patients as part of the adapted-dose group (we did not model patients who expressed CYP3A5 in the adapted-dose group because there were too few to provide robust models). Again, these results supported the involvement of a complex gene network related to interindividual variability in the dose requirement for Tac. Notably, because the study population was 89% white, it cannot be assumed that the predictions can be generalized to patients from other ethnic groups.

Weight and significance of individual genes in the predictive models

The SNPs presented in Figure 2 are contributors to predictive models, and their contribution must be considered when taking into account their interrelation with other variants included in the model. To identify new potential candidate genes that would be relevant for Tac metabolism at the individual level, we analyzed the behavior of the individual genes in the PLS1 model performed 2 mo after transplantation (i.e. day 60). We chose this model because its performance was the best, with a reasonable complexity level, and Tac doses and C_0 were relatively stable at this time (Table 1). The contribution of each SNP in the model (i.e. model weight) was highly variable, with cytochrome P450 gene variants having important contributions overall (Figure 4). Looking at the individual contribution of each SNP to Tac dose responses computed statistically with ANOVA and the linear mixed model, we found that the variants of cytochrome b5 reductase 3 (*CYB5R3*), carnosine dipeptidase 1, tumor protein 73, solute carrier family 28A3 (*SLC28A3*), cyclin-dependent kinase 8, *CYP3A4, CYP3A5*, and *CYP4F2* may individually explain part of the variance of the Tac pharmacokinetic parameters, with significant p-values for ANOVA (≤0.001) and the linear mixed model (≤0.05) (Figure 4), independent of their interaction with other genes in the models. All of these candidate genes were confirmed when the analysis was performed in the white population except for *CYB5R3*; this result might be a consequence of differences in minor allele frequencies between white and nonwhite participants.

Identification and validation of SLC28A3 as a candidate gene

Among these genes, we focused on the SNP rs10868152, which is an intronic variant of *SLC28A3* corresponding to a C/T substitution at the position 843555628 of chromosome 9. SLC28A3 is a transporter, and variants have been associated with variations in drug disposition (21,22). The distribution of the log (Tac C_0/dose) values across the three genotypes was noteworthy (Figure 5B). Indeed, the distribution of the log (Tac C_0/dose) values of patients with the TT genotype at day 60 after transplantation was skewed toward lower values, indicating that TT carriers may require higher Tac doses to obtain a therapeutic log (Tac C_0/dose) (Figure 5A) and arguing for a recessive effect (compare with *CYP3A5*). Consistently, the distribution of the log (Tac C_0/dose) at each time point after transplantation (day 10, 14, 30, 60, and 90) suggests that carriers of the TT genotype have systematically lower log (Tac C_0/dose) compared with CT and CC carriers (Figure 5B). These differences in distribution were significant ($p \leq 0.05$) with parametric tests (repeated-measures ANOVA) and with the linear mixed model ($p \leq 0.001$). On the basis of the SNPs present in the microarray, rs10868152 was not found in linkage disequilibrium with other variants of genes involved in drug metabolism (in contrast to *CYP3A5)* (Figure 5C). In addition, the 1000 Genomes Project and HapMap genotype databases indicate that rs10868152 was not in linkage disequilibrium with other genes in black and white populations.

To confirm the role of this variant in Tac metabolism, we tested the statistical difference between the distributions of the log (Tac C_0/dose) for the three possible genotypes in an independent cohort of 189 KTRs from the Necker Hospital under Tac treatment and found that Tac C_0/dose was significantly lower in this group compared with carriers of the CT or CC genotype (Figure 6). *SLC28A3* variants were not associated with graft function, histology, and survival (data not shown). Together, these results indicate that the rs10868152 variant of *SLC28A3* is associated with increased Tac bioavailability, the mechanism of which remains to be elucidated.

Discussion

Using a high-throughput genetic screening approach to predict variability of Tac dose requirement in KTRs, we demonstrated (i) that SNP networks explain 30–70% of the interpatient variability of Tac metabolism, depending on the model generated and the time after transplantation; (ii) that gene interaction networks related to oxidoreductase functions and monooxygenase activity, including *CYP3A4* and *CYP3A5,* have a major impact on Tac metabolism; and (iii) that the multidrug transporter ABCC8 and the nucleoside carrier SLC28A3 appear to be involved in Tac metabolism.

Our results indicate that nearly 70% of the dose response to Tac can be predicted by the combination of gene variants using a multivariate learning model generated later after transplantation (days 60 and 90), probably because Tac C_0 is more stable with less dose variation at later times than at early posttransplantation periods. This suggests that reduced background noise related to less variance in Tac C_0 allows for the identification of more complex (i.e. enriched) gene networks involved in Tac metabolism. This idea was reinforced by the results obtained for the two subgroups of patients (control group receiving a fixed dose and *CYP3A5*3/*3* patients as part of the adapted-dose group) at day 10 (Table S3A). By

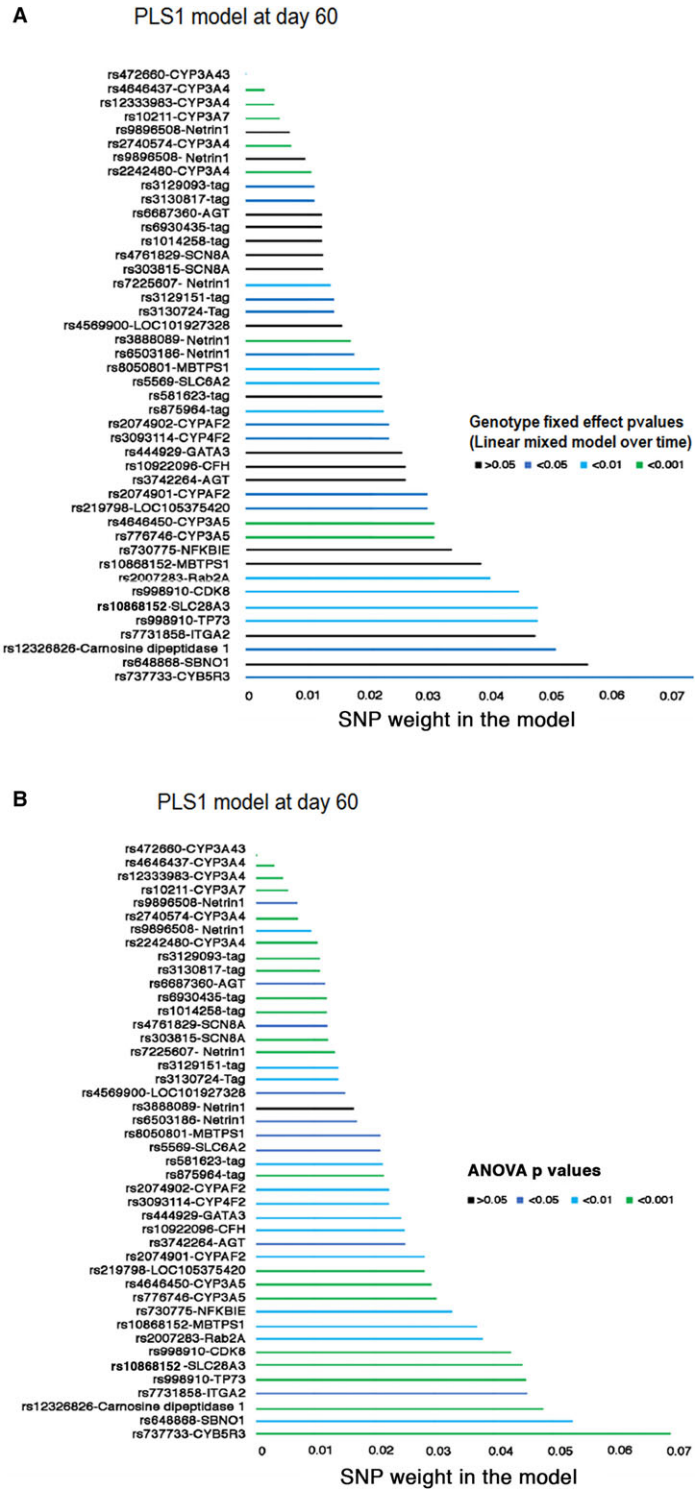

Figure 4: Effects of model weight and statistical genotype effect for each SNP on log (Tac C_0/dose) within the predictive model 60 days after transplantation. The line associated with each SNP is characterized by its length corresponding to the SNP's weight affected by the model and a color corresponding to the significance level (i.e. p-value threshold) of the statistical genotype effect of the SNP on log (Tac C_0/dose) assessed with (A) repeated measures (ANOVA) and (B) a linear mixed model. ANOVA, analysis of variance; log (Tac C_0/dose), logarithmically transformed tacrolimus blood concentration per dose; PLS, partial least squares; SNP, single-nucleotide polymorphism.

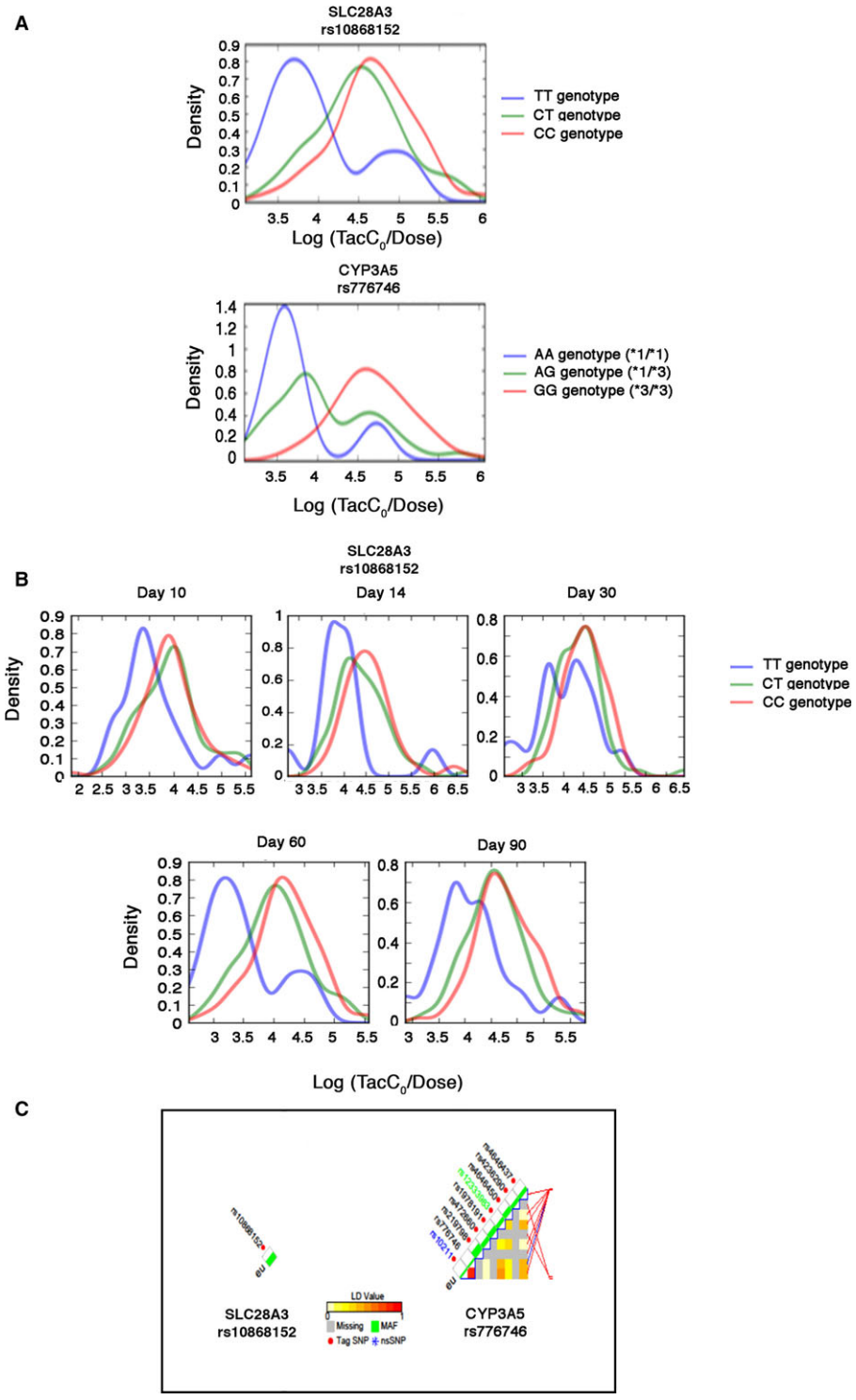

Figure 5: The rs10868152 gene variant of *SLC28A3* affects tacrolimus dose requirement. (A) Distribution of the log (Tac C_0/dose) values according to the three genotypes of *SLC28A3* and *CYP3A5* (serving as a control) at day 60 after transplantation. (B) Distribution of the log (Tac C_0/dose) values according to the three genotypes of *SLC28A3* at each period of time after transplantation. (C) Linkage disequilibrium analysis for rs10868152 (*SLC28A3*) and rs774746 (*CYP3A5,* serving as control) taking into account all of the other SNPs present on the microarray. log (Tac C_0/dose), logarithmically transformed tacrolimus blood concentration per dose; SNP, single-nucleotide polymorphism; LD, linkage disequilibrium; MAF, minor allelic frequency; nsSNP, nonsynonymous single nucleotide polymorphism.

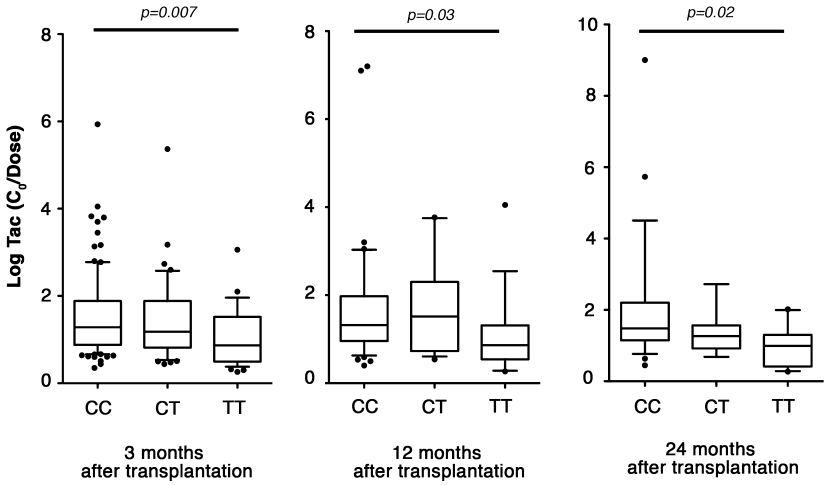

Figure 6: Validation of the association between rs10868152 and tacrolimus dose requirement. Box and whisker plots representing the distribution of tacrolimus daily doses 3, 12, and 24 mo after transplantation in the Necker cohort. The p-values were calculated using a Kruskal–Wallis one-way analysis of variance. log (Tac C_0/dose), logarithmically transformed tacrolimus blood concentration per dose.

fixing the initial dose at day 7, the resulting models achieved much higher predictive scores than the model including all initial patients. We observed much better predictions of Tac exposure on day 10, when the models were developed for each group separately. We can assume that by fixing different doses for three groups of patients at day 7 (initially defined for the objectives of the primary study), additional variability caused by the initial experimental design might have been introduced and likely altered the prediction of the models developed for all groups combined. Because of this additional background noise, our model at day 10 may have had more difficulties in capturing Tac variability related to patient phenotype than our models for subgroups with similar drug dosage. When considering the response over all follow-up periods, the model including all patients explained up to 30% of the interindividual variability in the pharmacokinetic response to Tac with gene variants, most of which were integrated in a functional network related to drug metabolism and transport, with *CYP3A4* and *CYP3A5* playing a central role. *CYP3A4* and *CYP3A5* were the only genes that our study and previous studies using a customized SNP chip commonly identified as predictors of the response to Tac (23).

From a methodological standpoint, the efficiency of our approach confirmed the conclusions of several studies about the use of an ensemble variable-selection strategy based on univariate filters and combined with multivariate modeling (9,10). Given the high dimension of the genomic data compared with the low number of observations and the high multicolinearity among the variables (SNPs), predictive algorithms usually integrate both dimension reduction and multivariate modeling, and factors such as variable selection for dimension reduction appear to be critical for the performance of the algorithm (19,24,25). Along this line, univariate variable selection appears to provide the best results in terms of accuracy, stability and interpretability of the genetic signatures compared with multivariate variable selection (26). Compared with multivariate variable selection, univariate filters proved to be efficient in reducing the overfitting risk; building more generalizable, robust and interpretable models; and highlighting the relevant biological processes underlying drug responses (27,28). We reached the same conclusion by testing and comparing our data using univariate feature selection based on the Fisher exact test and MI with the recursive feature-elimination multivariate method. Nevertheless, single-variable selection technique may be sensitive to small perturbations in the training data, and this effect is largely enhanced by the low ratio of samples to variables, leading to unstable signatures. Moreover, given the wide variety of possible variables for selection, and according to the "no free lunch" theorem (27,29), little overlap between signatures at the gene level captured by the different learning processes is generally observed. To overcome this phenomenon, ensemble variable-selection techniques that combine multiple sets of selected variables estimated based on random subsamples have been used increasingly, improving the stability of the variable selection process (9,10). Although we did not achieve significant performance using our two single-feature selection techniques alone (data not shown), our ensemble feature-selection scheme did. A possible improvement of our approach would be the use of nonlinear kernel-based PLS regression, which has been shown to be more effective in defining the correlation between drug response and gene expression (19).

We provided evidence of a role for the *SLC28A3* gene variant rs10868152 in the modulation of Tac pharmacokinetic parameters. Importantly, this association was validated in an independent cohort of KTRs, and the variant is not in linkage disequilibrium with either the other genes variants in the chip or the 1000 Genomes Project and HapMap genotype databases. Moreover, SLC28A3 has broad specificity for pyrimidine and purine nucleosides, and variants of the gene have been associated with variation in the responses to ribavirin, gemcitabine and anthracyclines (21,22). It is tempting to speculate that SLC28A3 and ABCC8 can transport Tac; however, the transport of Tac by SLC28A3 by a direct mechanism remains to be elucidated by functional studies. The rs2237991 variant of *ABCC8*, a member of the multidrug resistance subfamily that also modulates the ATP-sensitive K channel, has been captured by the predictive model of log (Tac C_0/dose) over all follow-up times after transplantation. These two gene variants suggest that transporters other than MDR1/ABCB1 are likely implicated in Tac pharmacokinetics.

In conclusion, we produced a comprehensive analysis of the networks of interacting genes involved in Tac metabolism and found that up to 70% of the dose-requirement variability can be predicted by variants encoding metabolism enzymes and transporters. Further studies are needed to validate the clinical utility and impact of predictions tools, including pharmacogenetics, on transplant outcomes. Moreover, we provided evidence of the critical role of transporters including the multidrug transporter ABCC8 and the nucleoside transporter SLC28A3.

Disclosure

The authors of this manuscript have no conflicts of interest to disclose as described by the *American Journal of Transplantation*.

References

1. Ventola CL. The role of pharmacogenomic biomarkers in predicting and improving drug response: Part 2: Challenges impeding clinical implementation. P T 2014; 38: 624–627.
2. Ventola CL. Role of pharmacogenomic biomarkers in predicting and improving drug response: Part 1: The clinical significance of pharmacogenetic variants. P T 2013; 38: 545–560.
3. Maliepaard M, Nofziger C, Papaluca M, et al. Pharmacogenetics in the evaluation of new drugs: A multiregional regulatory perspective. Nat Rev Drug Discov 2013; 12: 103–115.
4. Wallemacq P, Armstrong VW, Brunet M, et al. Opportunities to optimize tacrolimus therapy in solid organ transplantation: Report of the European consensus conference. Ther Drug Monit 2009; 31: 139–152.
5. Elens L, Bouamar R, Hesselink DA, et al. A new functional CYP3A4 intron 6 polymorphism significantly affects tacrolimus pharmacokinetics in kidney transplant recipients. Clin Chem 2011; 57: 1574–1583.
6. Elens L, Bouamar R, Shuker N, Hesselink DA, van Gelder T, van Schaik RH. Clinical implementation of pharmacogenetics in kidney transplantation: Calcineurin inhibitors in the starting blocks. Br J Clin Pharmacol 2014; 77: 715–728.
7. Thervet E, Anglicheau D, King B, et al. Impact of cytochrome p450 3A5 genetic polymorphism on tacrolimus doses and concentration-to-dose ratio in renal transplant recipients. Transplantation 2003; 76: 1233–1235.
8. Thervet E, Legendre C, Beaune P, Anglicheau D. Cytochrome P450 3A polymorphisms and immunosuppressive drugs. Pharmacogenomics 2005; 6: 37–47.
9. Awada W, Khoshgoftaar TM, Dittman D, Wald R, Napolitano A. A review of the stability of feature selection techniques for bioinformatics data. In: IEEE, editor. Information Reuse and Integration (IRI). Las Vegas, NV: 2012 IEEE 13th International Conference, 2012; p. 356–363.
10. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 2010; 26: 392–398.
11. Davis CA, Gerick F, Hintermair V, et al. Reliable gene signatures for microarray classification: Assessment of stability and performance. Bioinformatics 2006; 22: 2356–2363.
12. Thervet E, Loriot MA, Barbier S, et al. Optimization of initial tacrolimus dose using pharmacogenetic testing. Clin Pharmacol Ther 2010; 87: 721–726.
13. LeGatt DF, Shalapay CE, Cheng SB. The EMIT 2000 tacrolimus assay: An application protocol for the Beckman Synchron LX20 PRO analyzer. Clin Biochem 2004; 37: 1022–1030.
14. Stram DO. Tag SNP selection for association studies. Genet Epidemiol 2004; 27: 365–374.
15. Principe JC, Xu D, Zhao Q, Fisher Iii JW. Learning from examples with information theoretic criteria. J VLSI Signal Process Syst Signal Image Video Technol 2000; 26: 61–77.
16. Torkkola K. Feature extraction by non parametric mutual information maximization. J Mach Learn Res 2003; 3: 1415–1438.
17. Goncalves LB, Macrini JLR. Renyi entropy and cauchy-schwartz mutual information applied to mifs-u variable selection algorithm: A comparative study. Pesqui Oper 2011; 31: 499–519.
18. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
19. Dasgupta N, Lin SM, Carin L. Modeling pharmacogenomics of the nci-60 anticancer data set: Utilizing kernel pls to correlate the microarray data to therapeutic responses. In: Springer, editor. Methods of microarray data analysis II. New York: Springer, 2002; p. 151–167.
20. Bai JP, Lesko LJ, Burckart GJ. Understanding the genetic basis for adverse drug effects: The calcineurin inhibitors. Pharmacotherapy 2010; 30: 195–209.
21. Lotsch J, Hofmann WP, Schlecker C, et al. Single and combined IL28B, ITPA and SLC28A3 host genetic markers modulating response to anti-hepatitis C therapy. Pharmacogenomics 2012; 12: 1729–1740.
22. Visscher H, Ross CJ, Rassekh SR, et al. Validation of variants in SLC28A3 and UGT1A6 as genetic markers predictive of anthracycline-induced cardiotoxicity in children. Pediatr Blood Cancer 2013; 60: 1375–1381.
23. Jacobson PA, Oetting WS, Brearley AM, et al. Novel polymorphisms associated with tacrolimus trough concentrations: Results from a multicenter kidney transplant consortium. Transplantation 2011; 91: 300–308.
24. Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012; 483: 603–607.

25. Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 2014; 32: 1202–1212.
26. Haury AC, Gestraud P, Vert JP. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS ONE 2011; 6: e28210.
27. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003; 3: 1157–1182.
28. Lai C, Reinders MJ, van't Veer LJ, Wessels LF. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics 2006; 7: 235.
29. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinformatics 2015; 2015: 198363.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Data S1: Supplementary methods.

Table S1: Demographic and clinical characteristics of the Tactique study cohort. Continuous variables are shown as median and interquartile range, and nominal variables are shown as number and proportion.

Table S2: Demographic and clinical characteristics of the Necker cohort study. Continuous variables are shown as median and interquartile range, and nominal variables are shown as number and proportion.

Table S3: Performance, significance and complexity of the predictive models at day 10 after transplantation (A, PLS1 model) and over all follow-up times (B, PLS2 model) applied to both the control group (fixed dose) and the *CYP3A5*3/*3* patients as part of the adapted-dose group.