Toxicol. Res.
Vol. 32, No. 4, pp. 289-300 (2016)
Open Access http://dx.doi.org/10.5487/TR.2016.32.4.289
pISSN: 1976-8257 eISSN: 2234-2753 **Original Article**

Toxicological Research
Official Journal of
Korean Society of Toxicology
Available Online at http://www.ToxicolRes.org

Prediction of Non-Genotoxic Carcinogenicity Based on Genetic Profiles of Short Term Exposure Assays

Luis Orlando Pérez[1], Rolando González-José[1] and Pilar Peral García[2]

[1]*Instituto Patagónico de Ciencias Sociales y Humanas (IPCSH), Centro Nacional Patagónico (CENPAT), Boulevard Brown 2915, Puerto Madryn, PC 9120, Provincia de Chubut, Argentina*
[2]*Instituto de Genética Veterinaria "Fernando Noel Dulout"-CONICET, Facultad de Ciencias Veterinarias, Universidad Nacional de La Plata, Calle 60 y 118 S/N, PC 1900, La Plata, Provincia de Buenos Aires, Argentina*

(Received March 21, 2016; Accepted June 22, 2016)

Non-genotoxic carcinogens are substances that induce tumorigenesis by non-mutagenic mechanisms and long term rodent bioassays are required to identify them. Recent studies have shown that transcription profiling can be applied to develop early identifiers for long term phenotypes. In this study, we used rat liver expression profiles from the NTP (National Toxicology Program, Research Triangle Park, USA) DrugMatrix Database to construct a gene classifier that can distinguish between non-genotoxic carcinogens and other chemicals. The model was based on short term exposure assays (3 days) and the training was limited to oxidative stressors, peroxisome proliferators and hormone modulators. Validation of the predictor was performed on independent toxicogenomic data (TG-GATEs, Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System, Osaka, Japan). To build our model we performed Random Forests together with a recursive elimination algorithm (VarSelRF). Gene set enrichment analysis was employed for functional interpretation. A total of 770 microarrays comprising 96 different compounds were analyzed and a predictor of 54 genes was built. Prediction accuracy was 0.85 in the training set, 0.87 in the test set and increased with increasing concentration in the validation set: 0.6 at low dose, 0.7 at medium doses and 0.81 at high doses. Pathway analysis revealed gene prominence of cellular respiration, energy production and lipoprotein metabolism. The biggest target of toxicogenomics is accurately predict the toxicity of unknown drugs. In this analysis, we presented a classifier that can predict non-genotoxic carcinogenicity by using short term exposure assays. In this approach, dose level is critical when evaluating chemicals at early time points.

Key words: Toxicogenomics, Non-genotoxic carcinogen, Random forest

INTRODUCTION

Carcinogens are a large group of substances, organic and inorganic, that are directly involved in causing cancer. According to their mode of action they can be categorized as either genotoxic (GTX) or not genotoxic (NGTX). The former induces specific mutations or chromosome aberrations through direct interaction with DNA, usually by formation of covalent bonds (1). Such alterations are detected by a battery of tests (Ames test, chromosomal aberration, micronucleus assays) that measures the integrity and the structure of the DNA. Non-genotoxic drugs, on the other hand, represents chemicals capable of producing tumorigenesis by some secondary mechanism not directly related to DNA damage (2). Their activities are so diverse, that it is easier to define the properties they lack rather than the properties they possess. In general they are chemicals that do not induce DNA repair, and are negative in *in vivo* and *in vitro* tests for mutagenicity.

Non-genotoxic carcinogens have a wide variety of mechanisms of cancer induction including receptor mediated endocrine modulation, non-receptor mediated endocrine modulation, regenerative proliferation, oxidative stress, xenobiotic receptor activation, peroxisome proliferation, induction of inflammatory response and/or gap junction intercellular

Correspondence to: Pérez Luis Orlando, Instituto Patagónico de Ciencias Sociales y Humanas (IPCSH), Centro Nacional Patagónico (CENPAT), Boulevard Brown 2915, Puerto Madryn, PC 9120, Provincia de Chubut, Argentina
E-mail: orlandoperez@cenpat-conicet.gob.ar

List of Abbreviations: GTX, Genotoxic; NGTX, Non-genotoxic; NTP, National Toxicology Program; TG-GATE, The Toxicogenomics Project Genomics Assisted Toxicity Evaluation system.

inhibition (3). Free radical production (particularly ROS) is a common sub-mechanism enhanced by several non-genotoxic carcinogens. Basically, cellular damage is promoted when the balance between pro and anti-oxidants is disturbed and the oxidants are not properly neutralized.

The diverse mechanisms of action, the tissue specificity and the lack of genotoxicity make non-genotoxic identification a challenging task. Rodent bioassays are considered the best available method for detecting such carcinogens. Risk assessment is done combining data from bioassays, epidemiological data, toxico-kinetic and disposition studies (3). The rationale behind this approach is that many of the drugs known to be carcinogens to humans are also carcinogens to animals. Classical studies in rats involve exposures for periods that range from 13 to 14 weeks. However, a proportion of chemicals are detected at the end of a 2 year period, making the animal chronic exposure assay elaborate and costly intensive.

A rapid and sensitive method for detecting hepatocarcinogenicity in drug screening is a long sought target. Control of gene transcription is the main regulatory mechanism of biological systems. Gene expression precedes protein synthesis, cell proliferation and ultimately pathological modifications. Therefore, it should be the most sensitive point to detect early changes (4). The aim of this analysis was to build a model that distinguishes non-genotoxic liver carcinogens by using expression profiles from short term exposure chemical treatments in rodents. Experimental data were obtained from the toxicogenomic database DrugMatrix™, The National Toxicology Program (U.S. Department of Health and Human Services) and The Toxicogenomics Project Genomics Assisted Toxicity Evaluation system (TG-GATEs) (5,6).

MATERIALS AND METHODS

Experimental design and compounds. To evaluate molecular profiles, public available data from the National

Table 1. Groups of chemicals for classification analysis

Drugs	Analysis set
DrugMatrix	
Non Genotoxic carcinogens (n = 9) Carbon Tetrachloride (CCL4), Methapyrilene (MP), Cyproterone Acetate (CPA) Phenobarbital (PBT), Fenofibrate (FF), Clofibrate (CFB), Bezafibrate (BF), Diethylstilbestrol (DES), Gemfibrozil (GFZ) Genotoxic Carcinogens and non-hepatocarcinogens (n=9) 2-Acetylaminofluorene (2-AAF), 3-Methylcholanthrene (MCA), Albendazole (ALB), Doxorubicin (DOX), Ibuprofen (IBF), 1-Naphthyl Isothiocyanate (ANIT), Methyl salicylate (MS), Amiodarone (AMI), Hydrazine (N2H4)	Training set
Non Genotoxic carcinogens (n = 4) Clofibric Acid (CA), Carbamazepine (CBZ), Ethinylestradiol (EE), 17-methyltestosterone (MET) Genotoxic carcinogens and non-hepatocarcinogens (n = 29) Clotrimazole (CLOT), Nimesulide (NIM), Naproxen (NAP), Dexamethasone (DXM), Diclofenac (DFNa), Fluphenazine (FP), Clomipramine (CMP), Erythromycin (ERM), Meloxicam (MLX), Stavudine (D4T), Promethazine (PMZ), Valproic acid (VPA), Allyl alcohol (AA), Troglitazone (TGZ), Methimazole (MTZ), 6-Mercaptopurine (MP), Pioglitazone (PGZ), Tamoxifen (TMX), Altretamine (HMM), Chlorambucil (CBC), Carmustine (BCNU), Aflatoxin b1 (AFB1), N-nitrosodiethylamine (NDEA), Raloxifene (RLX), Lomustine (LS), Safrole (SF), Mitomycin-c (MMC), Streptozotocin (STZ), Cytarabine (ara-C)	Test set
Unknown (n = 26) Aminoglutethimide (AG), Closantel (CLO), Tandutinib (MLN518), Clomiphene (CLM), Sulindac (SULIN), Progesterone (PR), Cinnarizine (CIN), nystatin (NYS), indomethacin (IM), catechol (CC), ketorolac (KET), Isoeugenol (IEUG), Leflunomide (LEF), Finasteride (FIN), Danazol (DZ), Salicylamide (SA), Chloroxylenol (CXL), Balsalazide (BZ), Crotamiton(CROT), Zileuton (ZL), Propylthiouracil (PTU), Rosiglitazone (RGZ), Carbimazole (CBZ), Ketoconazole (KET), Modafinil (MO), Simvastatin (SIM)	Unknown/
TGGATE	
Non Genotoxic carcinogens (n = 13) Methyltestosterone (MTS), Monocrotaline (MCT), Ethinylestradiol (EE), Fenofibrate (FFB), Methapyrilene (MP), Phenobarbital (PBT), Thioacetamide (TAA), Carbon tetrachloride (CCL4), Clofibrate (CFB), WY-14643 (WY), Gemfibrozil (GFZ), Carbamazepine (CBZ), Acetamide (AAA) Genotoxic carcinogens and non-hepatocarcinogens (n = 21) Tamoxifen (TMX), Lomustine (LS), Colchicine (COL), Carboplatin (CBP), Acetamidofluorene (AAF), Doxorubicin (DOX), Naphthyl isothiocyanate (ANIT), ketoconazole (KC), Tetracycline (TC), Erythromycin ethylsuccinate (EME), Caffeine (CAF), Tannic acid (TAN), Promethazine (PMZ), Nimesulide (NIM), Ethanol (ETN), Gentamicin (GMC), Acetaminophen (APAP), Amiodarone (AM), Aspirin (ASA), Diclofenac (DFNa), Allyl alcohol (AA)	Validation set

Toxicological Program (NTP) was selected (GEO Accession number GSE57822). This entity performs pre-chronic and two year studies in laboratory animals in order to assess specific needs in toxicology, yielding the largest molecular toxicology reference. Briefly, arrays corresponding to 77 chemicals and their respective controls were downloaded from DrugMatrix (Table 1). Total data points were 363: three repeats per treatment involved 231 arrays and every treatment had 4 controls, in total 132 control arrays in 22 control groups. The Carcinogenic Potency Database was used as a first option to label the chemicals (7). Each array was obtained from test-compound treated and vehicle con-

Table 2. Class discrimination analysis by Random Forest in DrugMatrix data

Chemicals	Dose level (mg/mL)	Random Forest classification[a]
Training set (N = 18)		Predicted class
NGTX		
Carbon Tetrachloride (CCL4)	1175	0(*)
Methapyrilene (MT)	100	NGTXC
Cyproterone Acetate (CPA)	2500	NGTXC
Phenobarbital (PBT)	54	NGTXC
Fenofibrate (FF)	215	NGTXC
Clofibrate (CFB)	130	NGTXC
Bezafibrate (BF)	617	NGTXC
Diethylstilbestrol (DES)	280	NGTXC
Gemfibrozil (GFB)	700	NGTXC
GTX		
2-Acetylaminofluorene (2-AAF)	30	0
3-Methylcholanthrene (MCA)	300	0
Doxorubicin (DOXO)	3	0
Hydrazine (N2H4)	45	0
Non hepathocarcinogen		
1-Naphthyl Isothiocyanate (ANIT)	60	0
Methyl salicylate (MS)	444	0
Albendazole (ALB)	62	0
Amiodarone (AMI)	147	0
Ibuprofen (IBF)	263	0
Test set (n = 33)		
NGTX		
Clofibric Acid (CA)	448	NGTX
Carbamazepine (CBZ)	490	NGTX
Ethinylestradiol (EE)	1480	NGTX
17-Methyltestosterone (MET)	2000	NGTX
GTX		
Aflatoxin b1 (AFB1)	0.3	0
Carmustine (BCNU)	16	0
Chlorambucil (CBC)	4.5	0
Cytarabine (ara-C)	487	0
Lomustine (LS)	8.75	0
Mitomycin-c (MMC)	1.7	0
N-nitrosodiethylamine (NDEA)	34	0
Raloxifene (RLX)	650	0
Safrole (SF)	488	0
Streptozotocin (STZ),	138	0
Tamoxifen (TMX)	64	0
Non hepathocarcinogen		
6-Mercaptopurine (MP)	25	0
Allyl alcohol (AA)	32	0
Altretamine (HMM)	40	0
Clomipramine (CMP)	115	0
Clotrimazole (CLOT)	89	NGTXC(*)
Dexamethasone (DXM)	150	NGTXC(*)
Diclofenac (DFNa)	10	0
Erythromycin (ERM)	1500	0
Fluphenazine (FP)	2.5	0
Meloxicam (MLX)	33	0
Methimazole (MTZ)	100	0
Naproxen (NAP)	10	0
Nimesulide (NIM)	162	0
Pioglitazone (PGZ)	1500	0
Promethazine (PMZ)	113	0
Stavudine (D4T)	1400	NGTXC(*)
Troglitazone (TGZ)	1200	0
Valproic acid (VPA)	1340	0
Unknown		
Aminoglutethimide (AG)	350	NGTXC
Balsalazide (BZ)	1100	0
Carbimazole (CBZ)	400	0
Catechol (CC)	195	0
Chloroxylenol (CXL)	1915	0
Cinnarizine (CIN)	750	0
Clomiphene (CLM)	250	0
Closantel (CLO)	22	0
Crotamiton(CROT)	750	NGTXC
Danazol (DZ)	2000	0
Finasteride (FIN)	800	NGTXC
Indomethacin (IM)	12	0
Isoeugenol (IEUG)	1560	0
Ketoconazole (KET)	227	0
Ketorolac (KET)	48	0
Leflunomide (LEF)	60	0
MLN518	212	0
Modafinil (MO)	325	NGTXC
Nystatin (NYS)	134	0
Progesterone (PR)	164	0
Propylthiouracil (PTU)	625	Und[b]
Rosiglitazone (RGZ)	1800	NGTXC
Salicylamide (SA)	1300	0
Simvastatin (SIM)	1200	NGTXC
Sulindac (SULIN)	132	0
Zileuton (ZL)	450	0

Abbreviations: NGTXC, Non-genotoxic Carcinogen; GTX, Genotoxic compound; NH, Non hepatocarcinogen; 0, Negative for NGTXC.
[a]Results based on the OOB classification.
[b]Undetermined.
*Misclassified.

Table 3. Random Forest classification of TG-GATE data according to dose level

	Low dose		Medium dose		High dose	
Samples	Conc. (mg/mL)	Predicted class	Conc. (mg/mL)	Predicted class	Conc. (mg/mL)	Predicted class
NGTXC						
Acetamide	300	0(*)	1000	0(*)	200	0(*)
Carbamazepine	30	0(*)	100	0(*)	300	NGTX
Carbon tetrachloride	30	0(*)	100	0(*)	300	0(*)
Clofibrate	30	0(*)	100	0(*)	300	NGTX
Ethinylestradiol	1	0(*)	3	NGTX	10	NGTX
Fenofibrate	10	0(*)	100	NGTX	100	NGTX
Gemfibrozil	30	NGTX	100	NGTX	300	NGTX
Methapyrilene	10	0(*)	30	0(*)	100	0(*)
Methyltestosterone	30	0(*)	100	0(*)	300	0(*)
Monocrotaline	3	0(*)	10	0(*)	30	0(*)
Phenobarbital	10	0(*)	30	0(*)	100	NGTX
Thioacetamide	4.5	0(*)	15	NGTX	45	NGTX
WY-14643	10	NGTX	30	NGTX	100	NGTX
GTX						
Acetamidofluorene	30	0	100	0	300	0
Carboplatin	1	0	3	0	10	0
Colchicine	0.5	0	1.5	0	5	0
Doxorubicin	0.1	0	0.3	0	1	0
Lomustine	0.6	0	2	0	6	0
Naphthyl isothiocyanate	1.5	0	5	0	15	0
Tamoxifen	6	0	20	0	60	0
Non-Hepatocarcinogen						
Acetaminophen	300	0	600	0	1000	0
Allyl alcohol	3	0	10	0	30	0
Amiodarone	20	0	60	0	200	0
Aspirin	45	0	150	0	450	0
Caffeine	10	0	30	0	100	0
Diclofenac	1	0	3	0	10	0
Erythromycin ethylsuccinate	100	0	300	0	1000	0
Ethanol	400	0	1200	0	4000	0
Gentamicin	10	0	30	0	100	0
Nimesulide	10	0	30	0	100	0
Promethazine	20	0	60	0	200	0
Tannic acid	100	0	300	0	1000	0
Tetracycline	100	0	300	0	1000	0

*Misclassified.

trol-treated male rats after 72 hr of exposure with daily dosing (Sprangle-Drawley, 6~8 weeks old). Liver tissues (medial lobe) from three rats per chemical was collected and submitted to array processing. More data on the original experiments can be found in (8). Concentrations selected for each compound are summarized in Table 2.

The validation set was extracted from The Toxicogenomics Project Genomics Assisted Toxicity Evaluation system (TG-GATEs) (ExpressArray E-MTAB-800), a large-scale database of transcriptomics and pathology data useful for predicting the toxicity of new chemical entities (6). We downloaded data from rats exposed daily for 4 days at three doses (low, middle and high). Four hundred and seven arrays corresponding to 34 chemicals and their corresponding controls were obtained. Table 1 shows selected drugs and Table 3 their dose levels.

Data pre-processing. Complete ".CEL" files were downloaded from the National Toxicological Program, Department of Health and Human Services (USA) and the National Bioscience Database Centre (National Bioscience Database Center, Tokyo, Japan). Files belonged to the Affymetrix Rat Genome 230 2.0 GeneChip Array (Affymetrix, Santa Clara, CA, USA). Preprocessing adjustments were performed with Expression Console (Affymetrix). Additional information and raw data from the public repositories can be found at https://ntp.niehs.nih.gov/drugmatrix/index.html and http://dbarchive.biosciencedbc. jp/en/open-tggates/download.html/.

Each array (CEL file) was preprocessed and background corrected, normalized and summarized using RMA (Robust Multiarray Average) using Expression Console (Affymetrix) and Bioconductor packages of the R software (Fred Hutchinson Cancer Research Center, Seattle, USA) (9). For probe filtering, unspecific selection was carried out according to the interquartile range (IQR) (cut off value according to the IQR density plot).

Genes that had more than 1.5 fold increase/decrease relative to controls were chosen for further analysis. Additionally, each drug treatment was compared with its respective control using *t* test statistic. Results were further corrected for multi-testing using the Benjamini & Yekutieli (2001) procedure for (conservative) control of the false discovery rate (FDR), with 0.05 as the significance level (10). Features that met both criteria (*t*-test and 1.5 fold change) were combined in a single list of differentially expressed genes that resulted in 3778 probes that underwent classification analysis.

Class discrimination.

Feature selection by random forest: In order to discriminate NGTX from other drugs (GTX and non-carcinogens) we divided the DrugMatrix data in a training set, a test set, and an unclassified set. Data from (three) replicates per treatment were treated individually, i.e. not combined. The training set consisted of 18 compounds where half of them were classical non-genotoxic carcinogens: oxidative stressors, peroxisome proliferators and hormone modulators. The test set consisted of 33 compounds, 4 NGTXC and 29 genotoxins or non-hepatocarcinogens. A third group was built with DrugMatrix data: those with no conclusive data about non-genotoxic hepatocarcinogenicity (n = 26). The resulting groups are described in Table 1.

We used Random Forest (RF) for classification. Its performance is comparable to other machine learning methods and combined with variable selection aggressively reduces the set of genes (11). In this approach, training and test sets are constructed internally and randomly, by iteratively partition of the dataset. Many decision trees are constructed (in this case 10,000 per RF). For the *kth* tree, a random vector θ_k is created, independent of the other generated vectors but with the same distribution, and a tree is grown casting a unit vote for the most popular class at input x (12).

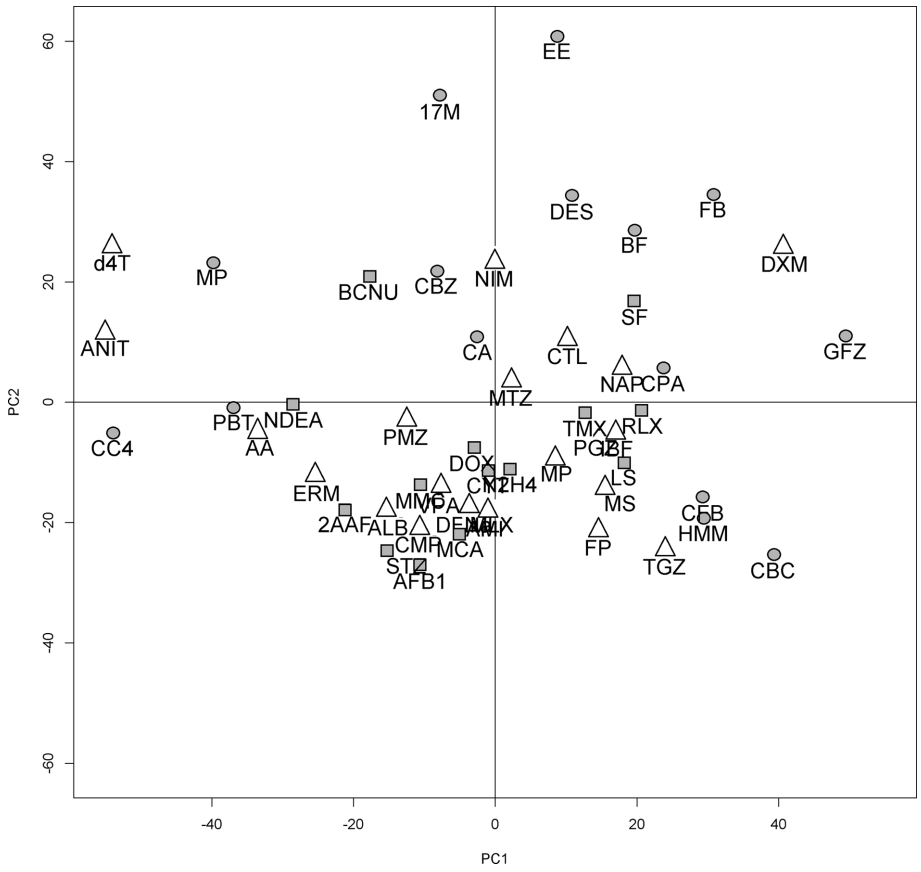

Fig. 1. Principal Component Analysis (PCA) of differential expressed genes for DrugMatrix chemicals. Each compound was averaged among replicates. Shapes indicate their class: circles correspond to non-genotoxic carcinogens, squares to genotoxins and triangles to non-hepatocarcinogens.

Table 4. High-scoring genes selected according to mean decrease accuracy

Affymetrix n°	*Symbol*	*Genename*
1367696_at	*Ifitm2*	*Interferon induced transmembrane protein 2*
1367780_at	*Pttg1*	*Pituitary tumor-transforming 1*
1368205_at	*Cfi*	*Complement factor I*
1368260_at	*Aurkb*	*Aurora kinase B*
1368742_at	*C5ar1*	*Complement component 5a receptor 1*
1368745_at	*Slc10a2*	*Solute carrier family 10 (sodium/bile acid cotransporter), member 2*
1368860_at	*Phlda1*	*Pleckstrin homology-like domain, family A, member 1*
1368991_at	*Smpd3*	*Sphingomyelin phosphodiesterase 3, neutral membrane*
1369031_at	*Il18bp*	*Interleukin 18 binding protein*
1369161_at	*Abcb4*	*ATP-binding cassette, subfamily B (MDR/TAP), member 4*
1369483_at	*Cd4*	*Cd4 molecule*
1370166_at	*Sdc2*	*Syndecan 2*
1370381_at	*Pnrc1*	*Proline-rich nuclear receptor coactivator 1*
1370828_at	*Zdhhc2*	*Zinc finger, DHHC-type containing 2*
1371170_a_at	*Il1a*	*Interleukin 1 alpha*
1371388_at	*Pdhb*	*Pyruvate dehydrogenase (lipoamide) beta*
1371577_at	*Ndufs1*	*NADH dehydrogenase (ubiquinone) Fe-S protein 1*
1371754_at	*Slc25a25*	*Solute carrier family 25 (mitochondrial carrier, phosphate carrier), member 25*
1371809_at	*Mrps18b*	*Mitochondrial ribosomal protein S18B*
1371893_at	*Col4a3bp*	*Collagen, type IV, alpha 3 (Goodpasture antigen) binding protein*
1371924_at	*Olfml3*	*Olfactomedin-like 3*
1372013_at	*Ifitm1*	*Interferon induced transmembrane protein 1*
1372044_at	*Tango2*	*Transport and golgi organization 2 homolog*
1372920_at	*Prodh*	*Proline dehydrogenase (oxidase) 1*
1374061_at	*Cd302*	*CD302 molecule*
1374537_at	*Chsy1*	*Chondroitin sulfate synthase 1*
1374540_at	*Cdca7*	*Cell division cycle associated 7*
1375861_at	*Nap1l5*	*Nucleosome assembly protein 1-like 5*
1376135_at	*Dars2*	*Aspartyl-tRNA synthetase 2 (mitochondrial)*
1377011_at	*Fry*	*Furry homolog (Drosophila)*
1377012_at	*Smarcad1*	*SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily a, containing DEAD/H box 1*
1377785_at	*Dhx40*	*DEAH (Asp-Glu-Ala-His) box polypeptide 40*
1379046_at	*Mlec*	*Malectin*
1379636_at	*Rmdn2*	*Regulator of microtubule dynamics 2*
1380066_at	*Tfr2*	*Transferrin receptor 2*
1381975_at	*Prune2*	*Prune homolog 2 (Drosophila)*
1382078_at	*Tlr8*	*Toll-like receptor 8*
1384240_at	*Agtr1a*	*Angiotensin II receptor, type 1a*
1385001_at	*Gsdmd*	*Gasdermin D*
1386080_at	*Hey1*	*Hes-related family bHLH transcription factor with YRPW motif 1*
1387029_at	*Cfh*	*Complement factor H*
1387243_at	*Cyp1a2*	*Cytochrome P450, family 1, subfamily a, polypeptide 2*
1387745_at	*Cd200r1*	*CD200 receptor 1*
1388301_at	*Uqcrc1*	*Ubiquinol-cytochrome c reductase core protein I*
1388460_at	*Capg*	*Capping protein (actin filament), gelsolin-like*
1389180_at	*Phkb*	*Phosphorylase kinase, beta*
1390426_at	*Notch1*	*Notch 1*
1390667_at	*Lrrc51*	*Leucine rich repeat containing 51*
1390839_at	*Pqlc3*	*PQ loop repeat containing 3*
1391269_at	*Pim2*	*Pim-2 proto-oncogene, serine/threonine kinase*
1392664_at	*Gpr182*	*G protein-coupled receptor 182*
1392990_at	*Sox17*	*SRY (sex determining region Y)-box 17*
1397317_at	*Itgb3*	*Integrin, beta 3*
1399030_at	*Wdr45*	*WD repeat domain 45*

Random Forest parameters were set to: *ntree* = 10,000; *nodesize* = 1; mtry = square root of number of genes. Once obtained the forest, the construction of the classifier was performed by a feature reduction algorithm, the backwards variable elimination (VarSel package), where the less important features are successively eliminated and out of the bag (OOB) error is continuously analyzed. The process fits random forests iteratively in the training set, at each step discarding the less important variables of previous models, but keeping the OOB error until it drops substantially (fraction dropped = 0.1). To evaluate stability of results and the prediction error rate, bootstrap (.632+ rule) was run through all the procedure (11,13). The reported error corresponded to samples not used to fit the random forest or perform feature reduction.

Pathway and gene analysis: Functional annotation based on Gene Ontology was tested while accounting for the topology of the GO graph. The methodology applied Fisher's exact test, which is based on gene counts. Each GO category was tested independently, searching for overrepresented terms within the group of differentially expressed genes (14).

RESULTS

Pre-processing. After normalization and log transforming the data, unspecific filtering was applied to the 31099 probes, leaving 10091 features. Differential expressed genes were identified for each treatment when compared with the set of corresponding controls by *t*-test and fold change. A unique DEG list from all treatments consisting of 3778 probes was built. The filtered DrugMatrix set was used for Principal Component Analysis (PCA). Fig. 1 shows a two-dimensional plot of the data. Each color represents a single compound defined by the rainbow pallet. Variance was relatively low for the first four components: PCA1 (12%), PCA2 (9%), PCA3 (8%), PCA4 (4%). Overall, there was no significant clustering among NGTX and other compounds, suggesting that further filtering steps were required to make a successful discrimination.

Feature selection by random forest. Random Forests analysis combined with a feature selecting algorithm was used to build the predictor. In the training set, the out of the bag error (OOB) for the initial random forest was 0.09

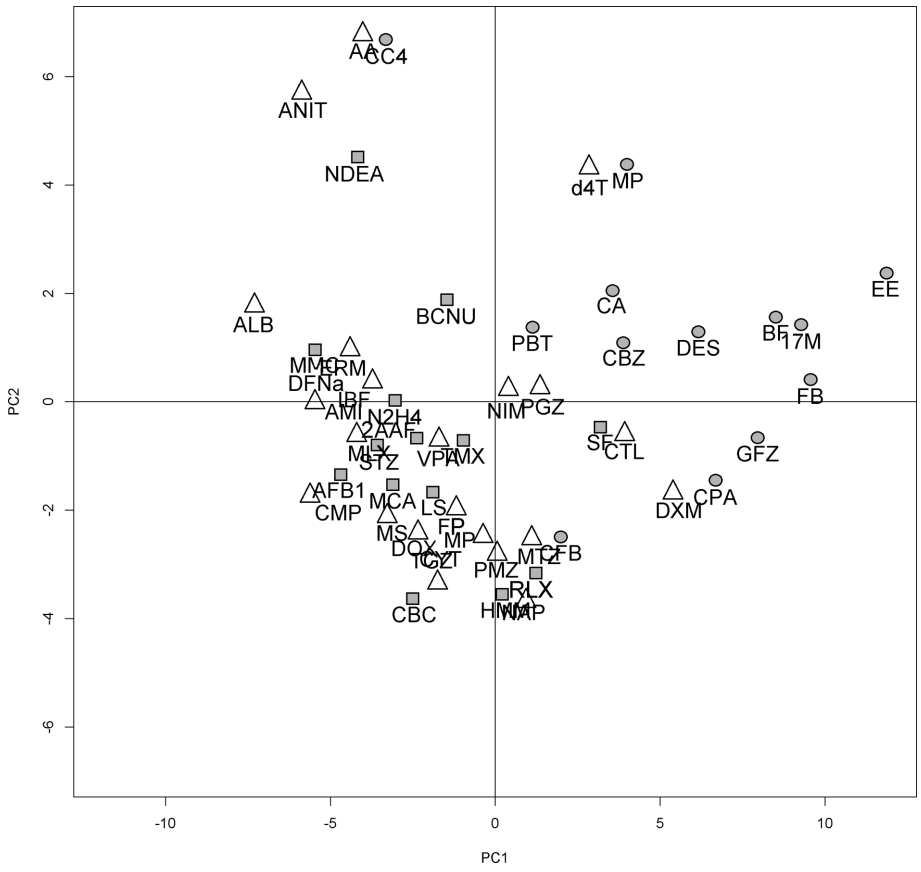

Fig. 2. Principal Component Analysis (PCA) performed with the 54 selected genes on the DrugMatrix set. Each compound was averaged among replicates. Shapes indicate their class: circles correspond to non-genotoxic carcinogens, squares to genotoxins and triangles to non-hepatocarcinogens. Note that non-genotoxic carcinogens (circles) clustered at the right of PC1, with the exception of CC4.

(81.5% sensitivity, 100% specificity). Fifty four variables were selected by variable selection without dropping the OOB error substantially (Table 4). Fig. 2 shows a Principal Component Analysis (PCA) constructed with the selected genes as variables. Each compound was averaged among replicates. The proportions of variance for the first three components were: PC1 39%, PC2 12% and PC3 9%. As can be seen on the plot, NGTX carcinogens were clustered at the right of the PC1 axis, with the exception of Carbon Tetrachloride. Fig. 3 represents the mean expression difference for each one of the predictor genes.

In order to assess an honest prediction of the error rate from the training data we performed bootstrapping (0.632) in which the random forest constructed for a certain number of variables was subsampled and compared. The prediction error rate among the bootstrap samples was 0.15 (Supplemental Fig. 1). Thirty genes were consistently selected (stability) in all the sub-samples (above 20%), with CYP1A2 being selected in 65%, Prodh in 32% and Itgb3 in 30%.

Prediction in the test set. The expression profiles of the test sample group were run through the obtained random forest. Overall, prediction error was lower than expected, 0.13. Sensitivity was 97% and specificity 81%. Three agents were misclassified as NGTX carcinogens: Clotrimazole, Stavudine and Dexamethasone. We found that six drugs had similar profiles to our predictor in the unclassified group: Aminoglutethimide, Crotamiton, Finasteride, Modafinil, Rosiglitazone and Simvastatin. Table 2 shows the results of the random forest in the DrugMatrix set.

Validation set. The performance of the classifier was tested in an independent dataset. We applied the random forest model to assess treated rat livers for 4 days and 34 drugs from the TG-GATE database. In contrast to DrugMatrix, three different doses were available at this time point

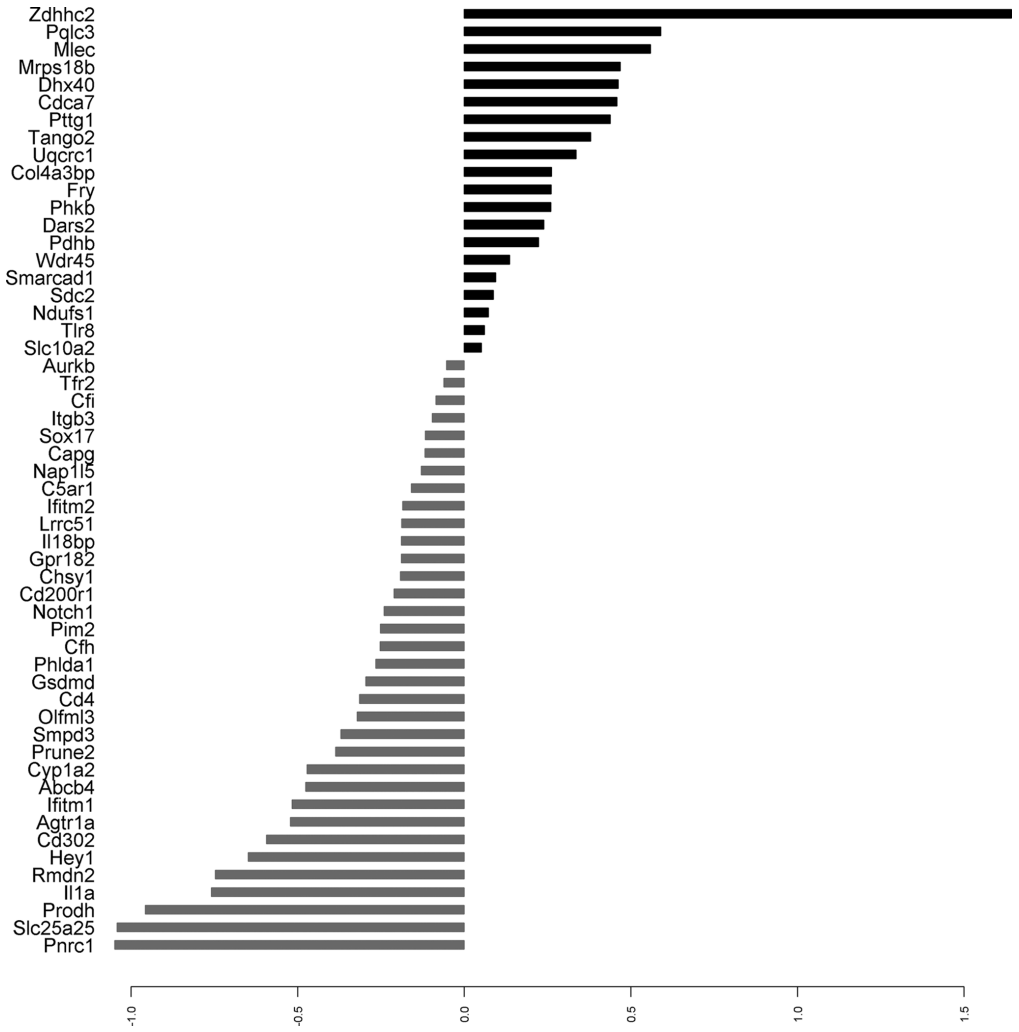

Fig. 3. Mean differential expression of the predictor genes in $\log_2$ scale. Black bars represent overexpressed genes; gray bars represent under-expressed genes.

Table 5. Top GO terms of enriched analysis according to Fisher's exact test

Biological process	Term	Annotated	Significant	Expected	F. classic
GO:0045333	Cellular respiration	28	5	0.42	4.8e-05
GO:0015980	Energy derivation by oxidation of organic compounds	50	6	0.75	8.3e-05
GO:0014823	Response to activity	23	4	0.35	0.00033
GO:0007507	Heart development	124	8	1.87	0.00043
GO:0006091	Generation of precursor metabolites and energy	68	6	1.03	0.00046
GO:0042157	Lipoprotein metabolic process	29	4	0.44	0.00082
GO:0051701	Interaction with host	30	4	0.45	0.00094
GO:0006869	Lipid transport	79	6	1.19	0.00104
Molecular function					
GO:0004872	Receptor activity	172	8	2.3	0.0016
GO:0008528	G-protein coupled peptide receptor activity	20	3	0.27	0.0022
GO:0001653	Peptide receptor activity	21	3	0.28	0.0025
GO:0060089	Molecular transducer activity	215	8	2.88	0.0066
GO:0001948	Glycoprotein binding	31	3	0.42	0.0077
GO:0038024	Cargo receptor activity	11	2	0.15	0.0089
GO:0005319	Lipid transporter activity	35	3	0.47	0.0109
GO:0004930	G-protein coupled receptor activity	49	3	0.66	0.0269

(low, middle and high). Prediction error varied significantly among dose levels. While all concentration levels yielded specificities of 100% (no false positives), sensitivity increased from 10% at low doses and 38% at medium doses, to 61% at high doses. Sensitivity could be increased at expense of specificity. Areas under the curve (AUC) for the ROC curve (Receiver Operating Characteristic) were 0.73 for low doses, 0.83 for middle doses and 0.87 for high doses. The results of the classification model applied to the validation set are described in Table 3.

Pathway analysis. In order to evaluate the biological pathways involved in the expression of the 54 highest scoring genes, we performed functional analysis of the GO terms that were significantly represented. Differential expressed genes were used as background list. We have found that significant genes were scattered among diverse GO Biological Process, most of them related to mitochondrial respiration, energy production and lipoprotein metabolism. Detoxification was located later in the list. Molecular functions included several terms of signal transducing and receptor activities. Table 5 shows the top eight GO annotations according to Fisher's exact test for BP (Biological Processes) and MF (Molecular Functions). There is a representation of how these terms are distributed over the GO graph in Supplemental Fig. 2.

DISCUSSION

Toxicogenomics is defined as the application of genomics technologies to study the adverse effects of pharmaceutical and environmental chemicals, with the hope of improving risk assessment and hazard screening (5). In this paradigm, public databases are essential tools for multiple actions, such as comparing profiles, discovering patterns or integrating networks. The largest databases available to date are the Japan Project Database (TG-GATE) and DrugMatrix (National Toxicological Program, NTP, USA) (5,6). Their diversity and standardized protocol make them the current reference. In recent years other databases were made available (CBES, CTD, etc.) and the challenge has been extended to data mining (15).

The mechanisms of non-genotoxic carcinogenicity are well described, although not well understood. It is assumed that most non-genotoxic carcinogens induce neoplasm and exhibit threshold tumor dose-responses. Classically, NGTX are regarded as tumor promoters but mechanisms such as regenerative hyperplasia, cytotoxicity or induction of oxidative damage are also key events for tumor initiation (16). There is still uncertainty whether NGTX are capable of initiate tumor events by themselves or whether they need coincidental factors. In this study, we focused our efforts to rodent hepatocarcinogens, a common target of safety assessment in drug development.

One of the most, if not the most, important step in class prediction is the correct assignment of the training set. New experimental data may change the status of a substance to another category limiting applicability of the predictor. Generalization of data is also critical, and in agreement with previous studies, false positives are often difficult to avoid (or "unavoidable"). In order to evaluate a complex process like carcinogenesis, we have selected compounds with three

modes of action to train our model: oxidative stressors, peroxisome proliferators and hormone modulators. Our hypothesis was that in some point, similarities among gene profiles would allow discrimination between non-genotoxic carcinogens and other class of chemicals.

We chose Random Forest as our machine learning method because is a robust classification algorithm. Feature reduction was performed by recursive variable elimination while maintaining class error (11). We report a predictor of 54 variables, a number that may be reduced to 24 without losing much prediction accuracy. After training the model, we applied our classifier to the DrugMatrix test set, which resulted in a relatively low total error. Only three false positives were detected and no false negatives among the OOB samples. The small sample size could explain why the prediction error of the test set was lower than the estimation of the training set (0.13). It is interesting that prediction accuracy increased with increasing doses in the validation set. This behavior could be explained by the fact that DrugMatrix doses for 3 day treatments were higher than TG-GATE in almost all compounds, frequently several times higher than the therapeutic dose.

One of the false positives detected was clotrimazole (CTL), an imidazole antimitotic. It is an inhibitor of p450, blocker of intracellular Ca^{++} stores and activator of the xenobiotic response (17). It has important correlation with our predictor, particularly in the induction of ZDHHC2 (Palmitoyltransferase ZDHHC2), a protein responsible for membrane binding and protein localization (receptors). Imidazole agents modulate p450 response by different gene subfamilies and display different actions (18). The other two miss-classified drugs were Dexamethasone (DXM) and Stavudine (d4T). When compared to our predictor, we found proximity in genes related to stress and inflammation (Cyp1a2, IL1A, COL4A3BP, SDC2, PTTG1, C5ar1, CFH, Cfi).

In line with previous studies, a mechanistic approach is a promising strategy for prediction as well as for pathway or functional category analysis. In this study, the most significant biological processes comprised those related to cellular respiration, energy derivation by oxidation of organic compounds and generation of precursor metabolites. This observation is in agreement with the known action of non-genotoxins and the role oxidative stress in carcinogenesis (3,19). Additionally, the presence of genes related to lipid metabolism is typical of profiles delivered by peroxisome proliferators. Others genes involved complement cascade, inflammation response, and some of them were related to cellular attachment.

Our signature included synthetic sex steroids (cyperone acetate, 17-methyltestosterone, diethylstilbestrol and ethinylestradiol). Several of them were shown to produce liver tumors in rats when given at therapeutic doses (20). Initially, the non-genotoxic effects of steroids were thought to be triggered by downstream genes of specific receptors, although increasing data suggest that they possess genotoxic action as well. We decided to keep them as NGTX carcinogens because the effect was complementary. Increased CYP activities and alterations in sterol metabolism are frequently associated with hepatomegaly in a non-genotoxic manner (3).

A steroid responsive gene, Interleukin-1a (IL-1a), is a high scoring gene in our signature. It is a critical cytokine whose expression is related with various aspects of human reproduction and expressed in a number of solid tumors. IL-1a serves as attractant by lymphocytes that keep the inflammation state that precedes malignancy. Sex steroid receptors downregulate and confine IL-1a expression, but its deregulation is a key inducer of proteolytic enzymes that degrade the extracellular matrix and remodel tissue (21).

Several studies have provided gene predictors for non-gentoxic hepato-carcinogens (22-28). Most of the methodologies used were based on support vector machines (SVM) coupled with a feature reduction algorithm. Published gene classifiers ranged from 9 to more than 100 probes and prediction accuracies were roughly equivalent. In general, data from rats under longer exposures provided higher accuracies. The composition of predictors differed among studies in length and diversity and only few genes overlapped. The reason for such heterogeneity is given by the differences in experimental designs. However, all these factors make biological interpretability more complex.

A recent study developed by Gusenleitner *et al*. (2014) scanned most of the NTP database and built a classifier that englobed liver carcinogenesis (combining both genotoxic and non-genotoxic carcinogens) (29). They validated their model with TG-GATE liver data and estimated liver carcinogenesis with an AUC of 0.78 (56.8% sensitivity and 82.91% specificity). It is interesting that our signature, although confined to a non-genotoxic subgroup of short term exposure, had in common CYP1a2 and ZDHHC2 as high scoring genes. Also, the presence of members of the superfamily of solute carrier proteins (SCL genes) in both signatures indicates the importance that membrane transporters have for drugs that "hitch-hike" one or another to enter the cells (30). On the other hand, the inclusion of genotoxic compounds to predict global carcinogenesis resulted in a signature with several cell cycle control genes, essential to the regulation of growth and apoptosis during mutagenic stress.

Toxicogenomics is a promising field with the hope of aid both the earlier elimination of toxic compounds in the drug pipeline and the discovery of new toxicity mechanisms. The biggest target is accurately predict the toxicity of unknown drugs. In this analysis, we presented a classifier that can predict non-genotoxic carcinogenicity by using short term exposure treatments from the NTP database. Although we are aware that the classifier does not have the prediction accuracy of signatures of long term exposure, early screen-

ing is an advantage that would allow prioritizing compounds for further testing.

ACKNOWLEDGMENTS

We would like to thank to M.Sc. Patricia Dell'Arciprete (National Patagonic Center-CONICET) for help in data processing.

REFERENCES

1. McGovern, T. and Jacobson-Kram, D. (2006) Regulation of genotoxic and carcinogenic impurities in drug substances and products. *TrAC, Trends Anal. Chem.*, **25**, 790-795.
2. Hayashi, Y. (1992) Overview of genotoxic carcinogens and non-genotoxic carcinogens. *Exp. Toxicol. Pathol*., **44**, 465-471.
3. Hernández, L.G., van Steeg, H., Luijten, M. and van Benthem, J. (2009) Mechanisms of non-genotoxic carcinogens and importance of a weight of evidence approach. *Mutat. Res*., **682**, 94-109.
4. Robinson, J.F., Pennings, J.L. and Piersma, A.H. (2012) A review of toxicogenomic approaches in developmental toxicology. *Methods Mol. Biol*., **889**, 347-371.
5. National Research Council (US) Committee on Applications of Toxicogenomic Technologies to Predictive Toxicology (2007) Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment, National Academies Press (US), Washington.
6. Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T. and Yamada, H. (2015) Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res*., **43**, D921-D927.
7. Fitzpatrick, R.B. (2008) CPDB: Carcinogenic Potency Database. *Med. Ref. Serv. Q.*, **27**, 303-311.
8. Ganter, B., Tugendreich, S., Pearson, C.I., Ayanoglu, E., Baumhueter, S., Bostian, K.A., Brady, L., Browne, L.J., Calvin, J.T., Day, G.J., Breckenridge, N., Dunlea, S., Eynon, B.P., Furness, L.M., Ferng, J., Fielden, M.R., Fujimoto, S.Y., Gong, L., Hu, C., Idury, R., Judo, M.S., Kolaja, K.L., Lee, M.D., McSorley, C., Minor, J.M., Nair, R.V., Natsoulis, G., Nguyen, P., Nicholson, S.M., Pham, H., Roter, A.H., Sun, D., Tan, S., Thode, S., Tolley, A.M., Vladimirova, A., Yang, J., Zhou, Z. and Jarnagin, K. (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J Biotechnol*., **119**, 219-244.
9. R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
10. Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist*., **29**, 1165-1188.
11. Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.
12. Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5-32.
13. Efron, B. and Tibshirani, R. (1997) Improvements on cross-validation: the 632+ bootstrap method. *J. Am. Stat. Assoc.*, **92**, 548-560.
14. Alexa, A., Rahnenfuhrer, J. and Lengauer, T. (2006) Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, **22**, 1600-1607.
15. Engelberg, A. (2004) Iconix Pharmaceuticals, Inc.-removing barriers to efficient drug discovery through chemogenomics. *Pharmacogenomics*, **5**, 741-744.
16. Melnick, R.L., Kohn, M.C. and Portier, C.J. (1996) Implications for risk assessment of suggested non-genotoxic mechanisms of chemical carcinogenesis. *Environ. Health Perspect.*, **104 Suppl 1**, 123-134.
17. Wu, S.N., Li, H.F., Jan, C.R. and Shen, A.Y. (1999) Inhibition of Ca2+-activated K+ current by clotrimazole in rat anterior pituitary GH3 cells. *Neuropharmacology*, **38**, 979-989.
18. Zhang, W., Ramamoorthy, Y., Kilicarslan, T., Nolte, H., Tyndale, R.F. and Sellers, E.M. (2002) Inhibition of cytochromes P450 by antifungal imidazole derivatives. *Drug Metab. Dispos.*, **30**, 314-318.
19. Gurer-Orhan, H., Orhan, H., Vermeulen, N.P. and Meerman, J.H. (2006) Screening the oxidative potential of several mono- and di-halogenated biphenyls and biphenyl ethers in rat hepatocytes. *Comb. Chem. High Throughput Screen*., **9**, 449-454.
20. El Etreby, M.F., Graf, K.J., Giinzel, P. and Neumann, F. (1979) Evaluation of effects of sexual steroids on the hypothalamic-pituitary system of animals and man *in Mechanism of toxic action on some target organs drugs and other substances. Proceedings of the european society of toxicology* (Chambers, P.L. and Giinzel, P. Ed.). Springer-Verlag, Berlin Heidelberg, pp. 11-40.
21. Singer, C.F., Kronsteiner, N., Hudelist, G., Marton, E., Walter, I., Kubista, M., Czerwenka, K., Schreiber, M., Seifert, M. and Kubista, E. (2003) Interleukin 1 system and sex steroid receptor expression in human breast cancer: interleukin 1alpha protein secretion is correlated with malignant phenotype. *Clin. Cancer Res*., **9**, 4877-4883.
22. van Delft, J.H., van Agen, E., van Breda, S.G., Herwijnen, M.H., Staal, Y.C. and Kleinjans, J.C. (2004) Discrimination of genotoxic from non-genotoxic carcinogens by gene expression profiling. *Carcinogenesis*, **25**, 1265-1276. [Erratum in: 2004, **25**, 2525; 2005, **26**, 511].
23. Nakayama, K., Kawano, Y., Kawakami, Y., Moriwaki, N., Sekijima, M., Otsuka, M., Yakabe, Y., Miyaura, H., Saito, K., Sumida, K. and Shirai, T. (2006) Differences in gene expression profiles in the liver between carcinogenic and non-carcinogenic isomers of compounds given to rats. *Toxicol. Appl. Pharmacol.*, **217**, 299-307.
24. Fielden, M.R., Brennan, R. and Gollub, J. (2007) A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by non-genotoxic chemicals. *Toxicol. Sci.*, **99**, 90-100.
25. Ellinger-Ziegelbauer, H., Gmuender, H., Bandenburg, A. and Ahr, H.J. (2008) Prediction of a carcinogenic potential of rat hepatocarcinogens using toxicogenomics analysis of short-term *in vivo* studies. *Mutat. Res.*, **637**, 23-39.
26. Auerbach, S.S., Shah, R.R., Mav, D., Smit, C.S., Walker, N.J., Vallant, M.K., Boorman, G.A. and Irwin, R.D. (2010) Predict-

ing the hepatocarcinogenic potential of alkenylbenzene flavoring agents using toxicogenomics and machine learning. *Toxicol. Appl. Pharmacol.*, **243**, 300-314.

27. Uehara, T., Minowa, Y., Morikawa, Y., Kondo, C., Maruyama, T., Kato, I., Nakatsu, N., Igarashi, Y., Ono, A., Hayashi, H., Mitsumori, K., Yamada, H., Ohno, Y. and Urushidani, T. (2011) Prediction model of potential hepatocarcinogenicity of rat hepatocarcinogens using a large-scale toxicogenomics database. *Toxicol. Appl. Pharmacol*., **255**, 297-306.
28. Lee, S.J., Yum, Y.N., Kim, S.C., Kim, Y., Lim, J., Lee, W.J., Koo, K.H., Kim, J.H., Kim, J.E., Lee, W.S., Sohn, S., Park, S.N., Park, J.H., Lee, J. and Kwon, S.W. (2013) Distinguishing between genotoxic and non-genotoxic hepatocarcinogens by gene expression profiling and bioinformatics pathway analysis. *Sci. Rep.*, **3**, 2783.
29. Gusenleitner, D., Auerbach, S.S., Melia, T., Gómez, H.F., Sherr, D.H. and Monti, S. (2014) Genomic models of short-term exposure accurately predict long-term chemical carcinogenicity and identify putative mechanisms of action. *PLoS ONE*, **9**, e102579.
30. He, L., Vasiliou, K. and Nebert, D.W. (2009) Analysis and update of the human solute carrier (SLC) gene superfamily. *Hum. Genomics*, **3**, 195-206.