TOXsIgN: a cross-species repository for toxicogenomic signatures

Thomas A. Darde, Pierre Gaudriault, Rémi Béranger, Clement Lancien, Annaelle Caillarec-Joly, Olivier Sallou, Nathalie Bonvallot, Cécile Chevrier, Séverine Mazaud-Guittot, Bernard Jégou, et al.

Databases and ontologies

TOXsIgN: a cross-species repository for toxicogenomic signatures

Thomas A. Darde[1,2], Pierre Gaudriault[1], Rémi Beranger[1], Clément Lancien[1], Annaëlle Caillarec-Joly[1], Olivier Sallou[2], Nathalie Bonvallot[1], Cécile Chevrier[1], Séverine Mazaud-Guittot[1], Bernard Jégou[1], Olivier Collin[2], Emmanuelle Becker[1], Antoine D. Rolland[1] and Frédéric Chalmel[1,*]

[1] Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, F-35000 Rennes, France, [2] Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA/INRIA) - GenOuest platform, Université de Rennes 1; F-35042 Rennes, France.

* To whom correspondence should be addressed.

Abstract

Motivation: At the same time that toxicologists express increasing concern about reproducibility in this field, the development of dedicated databases has already smoothed the path toward improving the storage and exchange of raw toxicogenomic data. Nevertheless, none provides access to analyzed and interpreted data as originally reported in scientific publications. Given the increasing demand for access to this information, we developed TOXsIgN, a repository for TOXicogenomic sIgNatures.

Results: The TOXsIgN repository provides a flexible environment that facilitates online submission, storage, and retrieval of toxicogenomic signatures by the scientific community. It currently hosts 754 projects that describe more than 450 distinct chemicals and their 8491 associated signatures. It also provides users with a working environment containing a powerful search engine as well as bioinformatics/biostatistics modules that enable signature comparisons or enrichment analyses.

Availability and Implementation: The TOXsIgN repository is freely accessible at http://toxsign.genouest.org. Website implemented in Python, JavaScript, and MongoDB, with all major browsers supported.

Contact: frederic.chalmel@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Humans today are exposed to numerous man-made environmental contaminants. By July 2017, over 130 million chemicals were listed in the Chemical Abstract Service (CAS), 120,000 of which are marketed in the European Union according to the European Chemicals Agency (ECHA) (https://echa.europa.eu/fr/information-on-chemicals). The potential toxicity of the overwhelming majority of these compounds remains almost uninvestigated (Tweedale, 2017).

Growing concerns about their potential adverse effects on human health and the environment led the European Commission to promulgate the regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) in 2007 (European Commission). This regulation has prompted innovative scientific programs to: *i)* screen for novel potential toxicants to which humans are exposed, especially, novel possibly carcinogenic, mutagenic, or reprotoxic (CPR) substances; *ii)* investigate the molecular mechanisms underlying their actions; and, finally, *iii)* develop predictive methods for assessing chemical hazards, ultimately intended to reduce the number of experimental tests on model organisms (RUSSELL and BURCH, 1959). Several landmark projects, such as Open TG-GATEs (Igarashi *et al.*, 2015), DrugMatrix (Ganter *et al.*, 2006), CMap (Lamb *et al.*, 2006), and a myriad of other high throughput studies, have been undertaken on the hypothesis that toxi-

cogenomic data, i.e., gene expression profiles in response to chemical exposure, will improve the prediction of their toxicity and the understanding of their mechanisms of action (Prathipati and Mizuguchi, 2016). This concept supplements the traditional ligand- and structure-based predictive approaches to assessing the safety of compounds by postulating that transcript profiling can effectively discriminate classes of compounds with similar adverse effects (Steiner *et al.*, 2004; Kavlock *et al.*, 2007).

Recently, the lack of reproducibility of biomedical research (Must try harder, 2012), including in the field of toxicology (Miller, 2014; George *et al.*, 2015; Poland *et al.*, 2014), has sparked apprehensions. Funding agencies, such as the National Institutes of Health (NIH), and other institutions share this concern and are participating in discussions of ways to enhance reproducibility in the environmental sciences (Collins and Tabak, 2014). Among the practical points to be considered when funding, planning and reporting toxicology studies, a crucial one is improving data transparency, including negative findings or contradictory results (Poland *et al.*, 2014). One effort in this context comes from the European Commission, which through the European Research Infrastructure Consortium (ERIC) is attempting to establish a model service for systems biology data management. Its objectives are to make biological data FAIR: Findable, Accessible, Interoperable, and Reusable (Wilkinson *et al.*, 2016). These goals are especially important in the field of toxicology. Resources such as CTD (Davis *et al.*, 2015) (http://ctdbase.org), diXa (Hendrickx *et al.*, 2015) (http://www.dixa-fp7.eu), ToxDB (Hardt *et al.*, 2016) (http://toxdb.molgen.mpg.de), CEBS (Lea *et al.*, 2017) (https://www.niehs.nih.gov/research/resources/databases/cebs/index.cfm) , NIH LINCS (Duan *et al.*, 2016) and Drug2Gene (Roider *et al.*, 2014) (http://www.drug2gene.com) have paved the way for improved storage, exchange, and analysis of toxicological data (Miller, 2015). The Toxicity Forecaster (ToxCast) developed by the US Environmental Protection Agency is another good example of an innovative tool designed to generate and share standardized data and predictive models for thousands of toxicants, including endocrine disruptors (Richard *et al.*, 2016).

Toxicogenomics is a scientific field that covers the acquisition, interpretation and storage of information about gene expression and associated protein activity to study the adverse effects of environmental and pharmaceutical chemicals on human health and the environment. Such studies use "–omics" technologies (for example, transcriptomics, proteomics, epigenomics, and related approaches) to discriminate molecular signatures, also called toxicogenomic signatures, that strongly correlate with genetic toxicity (Beedanagari *et al.*, 2014). It is now well established that investigators are supposed to submit their raw data to public databases such as the Gene Expression Omnibus (Barrett *et al.*, 2013) and ArrayExpress (Kolesnikov *et al.*, 2015). To the best of our knowledge, none of the existing toxicogenomics resources, including CTD and NIH LINCS, allows scientists to submit toxicogenomic signatures, i.e., the sets of genes showing altered status (in terms of gene expression, protein activity or epigenetic status) in individuals or their descendants after exposure to single or combined environmental factors. Although these signatures constitute the heart of studies in toxicogenomics, they usually only appear as supplementary tables and are accordingly complicated to reuse directly or compare with other data.

To meet the demands of scientists for easy access to such information (i.e., without the time-consuming downloading and re-processing of raw data), we developed TOXsIgN (for TOXicogenomic sIgNatures), a user-friendly resource that supports online submission, storage, and retrieval of toxicogenomic signatures. This repository is not intended either to archive raw data, as GEO and ArrayExpress (Kolesnikov *et al.*, 2015; Barrett *et al.*, 2013) do, or to replace existing toxicological databases (Davis *et al.*, 2015; Lea *et al.*, 2017; Hendrickx *et al.*, 2015; Duan *et al.*, 2016), but rather to complement these resources by acting as a distribution hub. One of the unique features of TOXsIgN is its ability to archive heterogeneous data and thus allows users to upload lists of overexpressed/underexpressed genes from different kinds of omics experiments (e.g., transcriptomic, proteomic, or epigenomic) and make them usable for cross-species and cross-technology comparisons. TOXsIgN is also intended to serve as a warehouse for toxicogenomics and predictive toxicology tools simultaneously based on and able to analyze the overall set of signatures deposited by the community. The TOXsIgN repository is freely accessible at https://toxsign.genouest.org.

2 Methods

2.1 Data storage, management, and retrieval

The database underlying TOXsIgN is based on MongoDB (https://www.mongodb.com), a free, open-source cross-platform document-oriented database program. This NoSQL database technology provides relevant features for TOXsIgN, such as flexible storage of massive and rapidly changing types of data, data replication, and JavaScript compatibility.

The TOXsIgN search engine is based on the implementation of an ElasticSearch server (https://www.elastic.co). Briefly, ElasticSearch is a NoSQL database manager with a powerful search engine primarily used to index textual data, which allows TOXsIgN to index all four layers of a TOXsIgN project information (the project, the studies, the assays, and the signatures) and simultaneously enables investigators to query the database according to several data categories, such as chemicals, genes, doses, species, cell lines, and tissues.

2.2 Web interface

The web interface of TOXsIgN was built with two open web frameworks, Pyramid (https://trypyramid.com/) and AngularJS (https://angularjs.org/). Pyramid embeds many features, such as a REST API, a JSON renderer, an SQLAlchemy Object-relational mapper (ORM), a Deform library to generate forms, and compatibility with SMTP servers. AngularJS, on the other hand, is an open JavaScript framework that extends traditional HTML vocabulary; it allows implementation of readable and quickly developable web environments.

To handle website traffic and provide data security, scalability, and deployment, all components of TOXsIgN (i.e., website server, MongoDB database, and Elasticsearch server) are hosted on separate individual virtual machines, with Docker (https://www.docker.com) as a container system. Briefly, Docker is an open-source virtualization system allowing easy deployment of container images, which in turn are lightweight, stand-alone, executable packages embedding everything needed to run a piece of software, such as code, runtime, system tools, libraries, and settings.

2.3 Workspace

The workspace supports the database and includes a job submission system to execute local scripts written in Python, R, or Tcl, and standalone programs. The corresponding web interface is based on the Pyramid framework as well as Python scripts and allows users to run modules and access their results. Execution of a module creates a new job from a

Python wrapper. The status of each job is available in the table accessible through the "running jobs" web page, which lists queued, running, and complete jobs coded respectively in gray, orange, and green.

Uploaded toxicogenomic signatures correspond to tab-delimited text files composed of Entrez Gene identifiers (IDs) (Maglott *et al.*, 2011). Before available modules are run from the workspace, these IDs are supplemented with related information, such as gene symbols, gene descriptions, taxonomy IDs, and HomoloGene IDs (NCBI Resource Coordinators, 2016) using "gene_info" and "homologene.data" files from the NCBI website (https://www.ncbi.nlm.nih.gov).

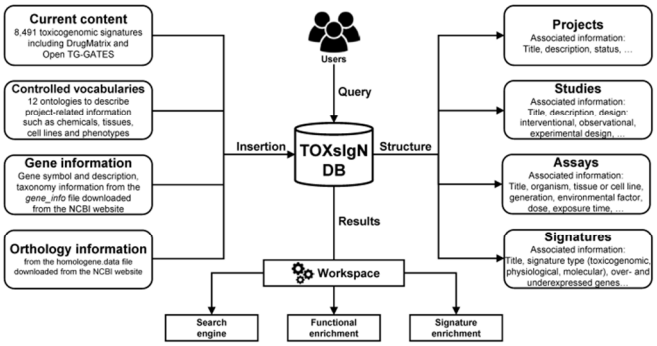

Fig. 1. Organization of the TOXsIgN database. The TOXsIgN database is organized in a four-layer architecture (Project > Study > Assay > Signature) associated with a unique identifier. The project layer covers one or several studies addressing specific questions. Each study in turn is associated with at least one assay that assesses the exposure of a given study model (e.g., cell culture, living animal, human population) to at least one chemical (e.g., pesticide, plasticizer, drug, or endocrine disruptor), physical (e.g., type of radiation or temperature) or biological (e.g., pathogen or parasite) agent at a given dose and for a given time of exposure. This organization makes TOXsIgN compatible with mixtures and transgenerational studies. At the time of writing, the database contains 8491 toxicogenomic signatures described by 12 different ontologies.

3 Results

3.1 TOXsIgN overview and current content

TOXsIgN information is organized in a four-layer architecture (Project > Study > Assay > Signature). Each layer is associated with a unique trackable identifier that can be reported in submitted manuscripts, like the raw data identifiers from GEO or ArrayExpress (Barrett *et al.*, 2013; Kolesnikov *et al.*, 2015) (Fig. 1). Scientists should thus submit **projects** (each of which receives an identifier with the "TSP" prefix), which are subsequently subdivided into **studies** (or subprojects, "TSE" prefix) addressing specific questions, and describing experimental **assays** ("TST" prefix) from which specific outcomes are extracted in the form of toxicogenomic **signatures** ("TSS" prefix), i.e., the set of genes positively or negatively affected in the corresponding assays. These assays can be performed directly in exposed individuals or in their descendants after exposure to a single substance or a combination of several. This definition underlines several unique features of TOXsIgN, specifically, its compatibility with: *i)* transgenerational studies; *ii)* chemical mixture studies – by default each assay is a mixture of at least one environmental factor; *iii)* a variety of environmental factors, including chemicals (e.g., pesticides, plasticizers, drugs, and endocrine disruptors), and eventually physical (e.g., radiations and temperature) and biological (e.g., pathogens and parasites) factors; and, *iv)* gene sets resulting from several kinds of (transcript-/prote-/epigen-)omics experiments. Importantly, TOXsIgN also allows scientists to submit and describe outcomes besides toxicogenomic signatures, such as physiological (e.g., association with a specific phenotype) and molecular (e.g., change in specific hormone concentrations) signatures for both interventional (participants undergo some kind of treatment so its impact can be evaluated) and/or observational studies (individuals for which different outcomes are measured) (Thiese, 2014). TOXsIgN thus collects a large number of heterogeneous signatures and simultaneously help to break down barriers between different fields in environmental sciences that remain boxed off from each other.

Currently, the TOXsIgN repository includes 754 projects for 911 transcriptomic studies of more than 450 compounds performed in humans, rats, mice, or drosophila (Supplementary Table 1 and Supplementary Information). Together these experimental assays correspond to 8491 toxicogenomic signatures, extracted from 32,688 microarray experiments that used 10 different technologies, including two major toxicogenomics resources, DrugMatrix and the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-GATES) (Ganter *et al.*, 2006; Igarashi *et al.*, 2015). DrugMatrix, set up by the National Institute of Environmental Sciences (NIES, USA) aimed at studying transcriptional responses in rats after exposure to 376 compounds in five different tissues (at multiple doses and multiple exposure times). On the other hand, Open TG-GATES is a collaborative project between the National Institute of Biomedical Innovation (NIB), the NIES and about 15 pharmaceutical companies that sought to study 150 chemicals and their transcriptional responses (at multiple doses and for multiple exposure times) in two rat tissues (liver and kidney). Our repository also hosts 326 physiological and molecular signatures from four interventional studies and four observational studies (Supplementary Table 1). In the near future, toxicogenomic signatures from other research programs will be included in TOXsIgN, such as CMAP, CEBS, diXa, and NIH LINCS (Lamb *et al.*, 2006; Lea *et al.*, 2017; Hendrickx *et al.*, 2015; Duan *et al.*, 2016).

3.2 Signature submission and access

In this quick and easy submission procedure, investigators will record all required information in a dedicated Excel template that embeds one tab for each layer (Project, Study, Assay, and Signature). This document integrates a dozen landmark controlled vocabularies allowing scientists, using ontologies as recommended (Hardy *et al.*, 2012; Smith *et al.*, 2007), to describe their toxicogenomic studies and their outcomes with precision (Supplementary Table 2). Once uploaded, the TOXsIgN webserver performs an initial evaluation of the Excel template to identify: *i)* "critical errors" about essential information that may not have been properly completed (such as the project title) and could therefore prevent the project upload; *ii)* "warnings" for important but not essential missing information (such as a PubMed identifier); and, *iii)* "information" for any other data not appropriately completed (such as additional information). If the system detects no "critical error", it next invites the user to upload the associated toxicogenomic signatures. Each signature comprises three one-column text files specifying: *i)* all interrogated genes; genes *ii)* positively (overexpressed for transcriptomic assays), and *iii)* negatively affected (underexpressed) in the corresponding assays. Because reliable and consistent identifier conversion is a complex problem, toxicogenomic signatures should be converted to Entrez Gene IDs from up-to-date resources (Huang *et al.*, 2008; Mudunuri *et al.*, 2009). For studies using Affymetrix GeneChip technologies, it is highly recommended that users normalize their raw data (CEL files) with the

Brainarray custom Chip Description Files (CDF) so that intensity values are not summarized for each probe set but directly for each Entrez Gene ID (Dai *et al.*, 2005) (*see* Supplementary Information).

By default, each submitted project and its related signatures are tagged with a "private" status, meaning that only authorized users (the owner but also coauthors) can access the uploaded data. At this stage, information can still be modified simply by uploading an updated version of the Excel template. A button is available on the web interface for each project to request the TOXsIgN administrators to switch it from private to public status. If "warnings" are still detected, this demand is rejected. The administrators will then help the investigators make the necessary modifications to activate the public status. Full instructions and examples of the submission procedure are provided in the website's tutorial section. The latest release of the Excel template document for uploading signatures in the repository and all ontologies (OBO files) used in TOXsIgN are available in a dedicated web page accessible through the "Download" tab on the main interface.

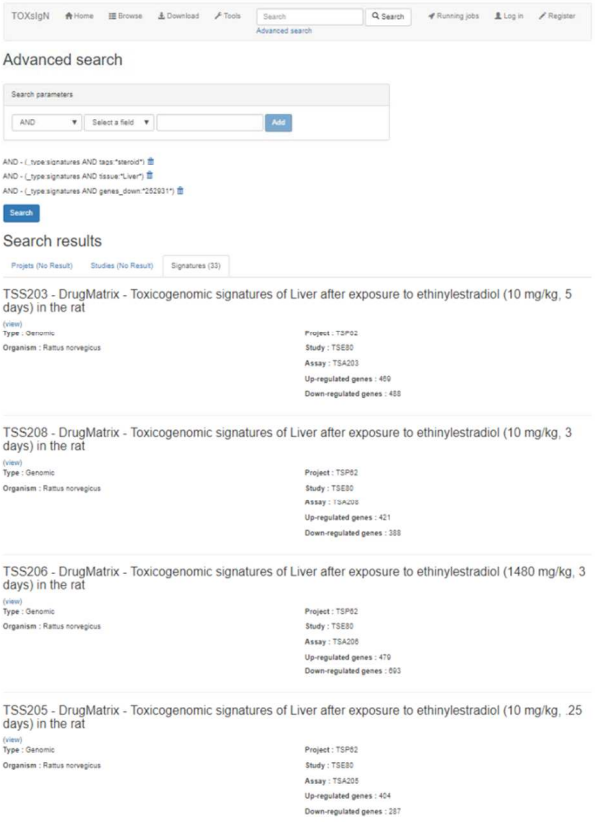

Fig. 2. The TOXsIgN advance search engine. Results for toxicogenomic signatures in which the expression of Cyp3a18 (Gene id: 252931) is downregulated in the rat liver after exposure to hydroxyl steroid compounds. From top to bottom, "search parameters", "full query", and "results table" are displayed. The "search parameters" box allows users to specify the details of their request from among projects, studies, assays, or signatures. For each request, the ElasticSearch query used to retrieve information is displayed on the top of the "results table". The "results table" is a three-tab panel displaying result sorted according to projects, studies, or signatures and providing a link to the corresponding page.

A powerful search engine is implemented to access all public information within TOXsIgN. Users can thus interrogate the database according to many distinct fields, such as environmental factors, organisms, tissues, and technologies. They can also easily make more advanced queries by using ontologies to describe toxicogenomic signatures. For instance, an investigator can retrieve 33 toxicogenomic signatures in which the expression of Cyp3a18 (Entrez Gene ID 252931, a gene encoding a member of the cytochrome P450 superfamily highly expressed in the liver (Nagata *et al.*, 1996)) is negatively affected in the liver after exposure to hydroxyl steroid compounds (Fig. 2). All scripts and data used are freely accessible in the TOXsIgN "Download" section.

3.3 Predicting adverse effects from toxicogenomic signatures

Predictive toxicology approaches based on toxicogenomic data seek to evaluate the toxicity of different compounds by using altered gene expression as an endpoint. In 2004, Steiner and colleagues established the proof-of-concept that similar toxicogenomic signatures imply similar adverse effects, using support vector machines (SVMs) to classify hepatotoxic and non-hepatotoxic chemicals based on transcriptomic data (Steiner *et al.*, 2004). Although the parameters may have been overfitted due to the small number of compounds, this approach correctly predicted hepatotoxic effects for 90% of known hepatotoxicants.

To illustrate the usefulness of archiving massive toxicogenomic signatures in a public repository we sought to demonstrate, but on a larger scale, the hypothesis formulated by Steiner and colleagues: that chemicals with similar toxicogenomic signatures share similar toxicity profiles (Supplementary Information). We used a subset of the current TOXsIgN repository that includes 3022 toxicogenomic signatures from transcriptomic assays from five rat tissues after exposure to 410 toxicants to evaluate the toxicological distance (concordance, i.e., the number of shared adverse effects between two toxicants) as well as the linear correlation between their toxicogenomic signatures (Pearson's correlation and Euclidean distance).

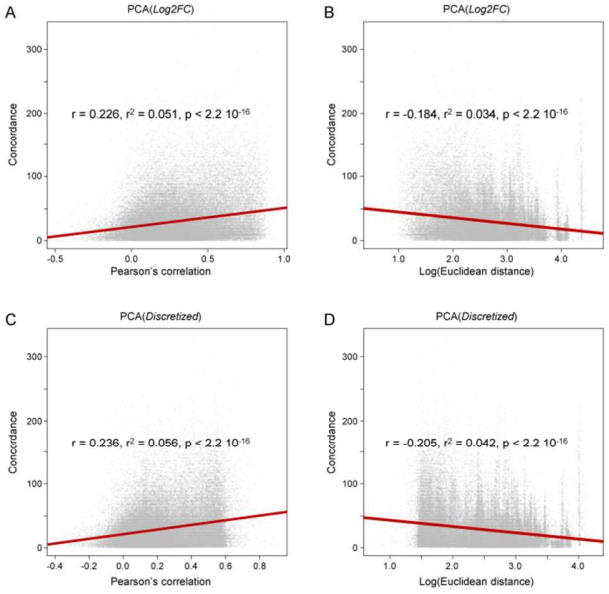

Fig. 3. Correlation between toxicogenomic and toxicological distances. The linear correlation between toxicogenomic distances (Pearson's correlation and Euclidean distance calculated between two toxicogenomic signatures) and toxicological distances (concordance, i.e., the number of adverse effects shared between two toxicants) was evaluated. Panels A and B show the association between toxicogenomic and toxicological distances, assessed by Pearson's correlation for the toxicogenomic distance on the log-fold-change matrix. Panels C and D show the degree of association with the use of a

discretized expression matrix, i.e., expression data for which fold-change information was discretized into only three distinct statuses (1, overexpressed after exposure; -1, underexpressed; and, 0, no differential expression).

We first showed a significant association between the toxicogenomic and toxicological distances ($r^2 = 0.051$, $P < 2.2\ 10^{-16}$) (Fig. 3, panels A-B). This result is in line with the hypothesis that a strong correlation between gene expression profiles implies similar toxicity characteristics. Interestingly, we also observed a linear correlation ($r^2 = 0.056$, $P < 2.2\ 10^{-16}$; panels C-D) when we discretized the expression data thanks to an expression matrix in which fold-change values information were simplified to only three distinct statuses: 1, genes overexpressed after exposure; -1, underexpressed; and, 0, no differential expression (*see* Supplementary Information). This finding supports the idea that toxicogenomic signatures can be archived as simplified text files, which substantially facilitates their submission, without excessively penalizing their predictive potential.

3.4 The signature enrichment analysis module for comparing toxicogenomic signatures

The core feature of the TOXsIgN workspace consists in the bioinformatics and biostatistics modules that allow retrieval, analysis, and comparison of the toxicogenomic signatures uploaded to the repository. The cross-species and cross-technology compatibility of this workspace relies on the conversion of toxicogenomic signatures into HomoloGene IDs before analysis by the different tools. The web page for each toxicogenomic signature includes a "Save in workspace" button that allows investigators to transfer it into the workspace from which all of the available modules can be accessed for further analysis.

In addition to the search engine and conversion tools, three other modules are currently available *via* the interface:

(1) *Signature comparison*. This module embeds an interactive Venn diagram viewer, called jvenn, for easy comparison of up to six selected toxicogenomic signatures (Bardou *et al.*, 2014).

(2) *Functional enrichment analysis*. This tool allows users to explore the mechanisms of toxicity of a given environmental factor by identifying the biological processes, molecular functions, cellular components (Ashburner *et al.*, 2000), and phenotypes (Human Phenotype Ontology web resource) (Köhler *et al.*, 2017) associated with its toxicogenomic signature. Briefly, it determines the significance of the resulting overlaps, according to the hypergeometric distribution.

(3) *Signature enrichment analysis*. This module seeks to identify toxicogenomic signatures in the repository closely related to the user's selected signature. Like the *Functional enrichment analysis* module, it uses the hypergeometric distribution to determine the significance of overlaps. It also includes a distance matrix calculation that uses either the Euclidean distance or Pearson's correlation to discriminate among the closely-related toxicogenomic signatures.

To illustrate the relevance of these modules, we uploaded into the workspace a toxicogenomic signature comprising 381 overexpressed and 494 underexpressed genes in rat livers after exposure to diethylstilbestrol (i.e., DES, a well-known estrogenic endocrine-disrupting chemical) (Korach and McLachlan, 1985; Li *et al.*, 2013) for 5 days, at a dose of 2.8 mg/kg (TOXsIgN signature identifier: TSS230). We then ran the *Signature enrichment analysis* module with its default parameters. It took about one minute to complete the job: consistently, three of the top-10 toxicogenomic signatures included experiments performed in rat livers after exposure to DES, but at different doses (three experimental conditions) (Fig. 4). The seven other signatures corresponded to other well-known estrogenic compounds sharing mechanisms of action similar to that of DES (such as ethinylestradiol, estriol, mestranol, and ß-estradiol) (MUECHLER and KOHLER, 1980; Simpson and Santen, 2015). This finding again confirms the hypothesis set forth by Steiner and colleagues (Steiner *et al.*, 2004).

Signature	r	R	n	N	Ratio	Pvalue	Adjusted Pvalue (Benjamini-Hochberg)	Zscore	Euclidean distance	Correlation distance
TSS203 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to ethinylestradiol (10 mg/kg, 5 days) in the rat	637	842	907	10770	57 %	0	0	73.1634644491449	21.7944947177034	0.729101755005883
TSS229 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to diethylstilbestrol (280 mg/kg, 3 days) in the rat	604	842	943	10770	51 %	0	0	67.336760881398	24.0208242989286	0.680260242000066
TSS226 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to diethylstilbestrol (280 mg/kg, 5 days) in the rat	649	842	1104	10770	50 %	0	0	66.5851006025112	25.4558441227157	0.674621819620628
TSS208 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to ethinylestradiol (10 mg/kg, 3 days) in the rat	542	842	770	10770	51 %	0	0	67.118009242705	22.9782508861521	0.67384691254637
TSS228 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to diethylstilbestrol (2.8 mg/kg, 3 days) in the rat	520	842	743	10770	49 %	0	0	65.4178016993892	23.3452350598575	0.66040973201856
TSS1179 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to beta-estradiol (150 mg/kg, 5 days) in the rat	565	842	1053	10770	42 %	0	0	58.3300260758659	27.6947648482525	0.604443584977887
TSS1183 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to beta-estradiol (150 mg/kg, 3 days) in the rat	449	842	721	10770	40 %	0	0	56.3862533358308	25.8263431402899	0.576911914376368
TSS783 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to mestranol (250 mg/kg, 5 days) in the rat	518	842	960	10770	40 %	0	0	55.795372188752	27.712812921102	0.576529708714113
TSS1117 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to estriol (313 mg/kg, 5 days) in the rat	556	842	1164	10770	38 %	0	0	53.7551493006238	29.8998327754521	0.563817014922414
TSS207 - DrugMatrix - Toxicogenomic signatures of Liver after exposure to ethinylestradiol (1480 mg/kg, 5 days) in the rat	620	842	1519	10770	36 %	0	0	51.688360069076	33.5111921802321	0.549224727537796

Fig. 4. Signature enrichment analysis for diethylstilbestrol. The toxicogenomic signature of diethylstilbestrol (DES) (2.8 mg/kg, 5 days, rat liver) was compared to all the toxicogenomic signatures indexed in TOXsIgN with the Signature enrichment analysis tool. N is the total number of HomoleGene IDs measured, R the total number of HomoleGene IDs meeting the criterion, n the total number of HomoleGene IDs in the selected signature, r the number of HomoleGene IDs meeting the criterion in this signature, and the Ratio corresponds to r/[union(n, R)]. The *P*-value and the adjusted *P-value* (Benjamini-Hochberg correction method) are obtained with the hypergeometric probability distribution. Euclidean distance and Pearson's correlation are also calculated to estimate the distance between each pair of toxicogenomic signatures.

4 Conclusion and perspectives

Our goal when designing TOXsIgN was to develop a new cross-species repository to allow scientists to submit the toxicogenomic signatures that they have published. Having been evaluated by experts during the peer-review process, these should be very high quality data. As pointed out in the *Nature* editorial "Must try harder" (Must try harder, 2012), it is essential to publish the results of well-conducted experiments, whether they are positive, negative, or uninterpretable (absent). We believe TOXsIgN constitutes a new alternative for toxicological data storage, accessibility, and especially reusability and should thus contribute to the transparency of toxicological experiments.

The success of this repository obviously depends on scientists' willingness to produce data that fit the FAIR criteria. Inversely, the success of the FAIR consortium also depends on the availability of repositories such as TOXsIgN. We think that making raw data available to the community may be considered an insufficient response, in view of the community's demand for direct and easy access to analyzed and interpreted data. We propose that the same effort should be made to encourage scientists to upload toxicogenomic and toxicological signatures in TOXsIgN before submission of manuscripts for publication. A win-win situation could thus be reached for all parts, since submitted studies would benefit from enhanced visibility, while the community would take

advantage of a continually updated user-friendly resource allowing TOXsIgN repository to expand.

On our side, we are currently integrating additional toxicogenomic signatures from other major toxicogenomics projects, such as CMAP, CEBS, diXa, and NIH LINCS (Lamb *et al.*, 2006; Lea *et al.*, 2017; Hendrickx *et al.*, 2015; Duan *et al.*, 2016). Likewise, we plan to make TOXsIgN compatible with other kinds of environmental agents, such as physical and biological factors. Altogether, we expect that this new resource can contribute significantly to risk assessment.

In addition to serving as a public repository, TOXsIgN is also intended to be a warehouse for toxicogenomic and predictive toxicology tools. Its modular design facilitates the implementation of additional bioinformatics modules relying on the deposited toxicogenomic signatures that will help investigators analyze and predict adverse effects of environmental factors relevant to their specific interests.

The TOXsIgN database will remain under constant development to offer more tools and enhance user experience. We are currently developing prediction and prioritization systems for chemical toxicity but also a module to extract toxicogenomic signatures automatically from raw data, in our own workflow. Another potential aspect is the addition of social features in TOXsIgN to enable several investigators to work on the same data. Finally, we plan to incorporate the ISA framework (http://isa-tools.org/), which provides rich descriptions of experimental metadata, to manage the TOXsIgN database and improve the FAIRness of available data. Together these efforts will improve TOXsIgN's utility, making it a front-line resource relevant to a large audience, including toxicologists, biologists, epidemiologists, and environmental scientists in general.

Acknowledgements

We thank the GenOuest bioinformatics facility for hosting the software as well as all members of the Research Institute for Environmental and Occupational Health (IRSET) for stimulating discussions. We also thank the Institut national de la santé et de la recherche médicale (Inserm), the Centre national de la recherche scientifique (CNRS), the Université de Rennes 1, and the French School of Public Health (EHESP) for supporting this work.

Funding

TOXsIgN is supported, built, and maintained by the Research Institute for Environmental and Occupational Health (IRSET), the French School of Public Health (EHESP), and the GenOuest BioInformatics core facility. This work was supported by the French agency for food and safety [ANSES n°EST-13-081 to F.C.]; the Fondation pour la recherche médicale [FRM n°DBI20131228558 to F.C.], and the European Union [FEDER to F.C]. Funding for open access charge: Institut national de la santé et de la recherche médicale (Inserm) and l'Université de Rennes 1.

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–9.

Bardou,P. *et al.* (2014) jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, **15**, 293.

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.*, **41**, D991-5.

Beedanagari,S. *et al.* (2014) Genotoxicity biomarkers. In, *Biomarkers in Toxicology*. Elsevier, pp. 729–742.

Collins,F.S. and Tabak,L.A. (2014) Policy: NIH plans to enhance reproducibility. *Nature*, **505**, 612–3.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.

Davis,A.P. *et al.* (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914-20.

Duan,Q. *et al.* (2016) L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *NPJ Syst. Biol. Appl.*, **2**, 16015.

European Commission REGULATION (EC) No 1907/2006 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/4.

Ganter,B. *et al.* (2006) Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database. *Pharmacogenomics*, **7**, 1025–44.

George,B.J. *et al.* (2015) Raising the bar for reproducible science at the U.S. Environmental Protection Agency Office of Research and Development. *Toxicol. Sci.*, **145**, 16–22.

Hardt,C. *et al.* (2016) ToxDB: pathway-level interpretation of drug-treatment data. *Database (Oxford).*, **2016**, baw052.

Hardy,B. *et al.* (2012) Toxicology ontology perspectives. *ALTEX*, **29**, 139–56.

Hendrickx,D.M. *et al.* (2015) diXa: a data infrastructure for chemical safety assessment. *Bioinformatics*, **31**, 1505–1507.

Huang,D.W. *et al.* (2008) DAVID gene ID conversion tool. *Bioinformation*, **2**, 428–30.

Igarashi,Y. *et al.* (2015) Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.*, **43**, D921-7.

Kavlock,R.J. *et al.* (2007) ToxCast TM : Developing predictive signatures for chemical toxicity.

Köhler,S. *et al.* (2017) The Human Phenotype Ontology in 2017. *Nucleic Acids Res.*, **45**, D865–D876.

Kolesnikov,N. *et al.* (2015) ArrayExpress update--simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113-6.

Korach,K.S. and McLachlan,J.A. (1985) The role of the estrogen receptor in diethylstilbestrol toxicity. *Arch. Toxicol. Suppl.*, **8**, 33–42.

Lamb,J. *et al.* (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science (80-.).*, **313**, 1929–1935.

Lea,I.A. *et al.* (2017) CEBS: a comprehensive annotated database of toxicological data. *Nucleic Acids Res.*, **45**, D964–D971.

Li,Y. *et al.* (2013) Diethylstilbestrol (DES)-Stimulated Hormonal Toxicity is Mediated by ERα Alteration of Target Gene Methylation Patterns and Epigenetic Modifiers (DNMT3A, MBD2, and HDAC2) in the Mouse Seminal Vesicle. *Environ. Health Perspect.*, **122**, 262–8.

Maglott,D. *et al.* (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52-7.

Miller,G.W. (2015) Data sharing in toxicology: beyond show and tell. *Toxicol. Sci.*, **143**, 3–5.

Miller,G.W. (2014) Improving Reproducibility in Toxicology. *Toxicol. Sci.*, **139**, 1–3.

Mudunuri,U. *et al.* (2009) bioDBnet: the biological database network. *Bioinformatics*, **25**, 555–6.

MUECHLER,E.K. and KOHLER,D. (1980) Properties of the Estrogen Receptor in the Human Oviduct and Its Interaction with Ethinylestradiol and Mestranol in Vitro*. *J. Clin. Endocrinol. Metab.*, **51**, 962–967.

Must try harder (2012) *Nature*, **483**, 509–509.

Nagata,K. *et al.* (1996) Isolation and characterization of a new rat P450 (CYP3A18) cDNA encoding P450(6)beta-2 catalyzing testosterone 6 beta- and 16 alpha-hydroxylations. *Pharmacogenetics*, **6**, 103–11.

NCBI Resource Coordinators (2016) ''''''. *Nucleic Acids Res.*, **44**, D7–D19.

Poland,C.A. *et al.* (2014) The elephant in the room: reproducibility in toxicology. *Part. Fibre Toxicol.*, **11**, 42.

Prathipati,P. and Mizuguchi,K. (2016) Systems Biology Approaches to a Rational Drug

Discovery Paradigm. *Curr. Top. Med. Chem.*, **16**, 1009–25.

Richard,A.M. *et al.* (2016) ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.*, **29**, 1225–1251.

Roider,H.G. *et al.* (2014) Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinformatics*, **15**, 68.

RUSSELL,W.M.S. and BURCH,R.L. (1959) The principles of humane experimental technique. *Princ. Hum. Exp. Tech.*

Simpson,E. and Santen,R.J. (2015) Celebrating 75 years of oestradiol. *J. Mol. Endocrinol.*, **55**, T1–T20.

Smith,B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–5.

Steiner,G. *et al.* (2004) Discriminating different classes of toxicants by transcript profiling. *Environ. Health Perspect.*, **112**, 1236–48.

Thiese,M.S. (2014) Observational and interventional study design types; an overview. *Biochem. medica*, **24**, 199–210.

Tweedale,A.C. (2017) The inadequacies of pre-market chemical risk assessment's toxicity studies-the implications. *J. Appl. Toxicol.*, **37**, 92–104.

Wilkinson,M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.