TOXsIgN a cross-species repository for toxicogenomic signatures


To cite this version:

Thomas A. Darde, Pierre Gaudriault, Rémi Béranger, Clement Lancien, Annaelle Caillarec-Joly, et al..
TOXsIgN a cross-species repository for toxicogenomic signatures. Bioinformatics, Oxford University Press (OUP), 2018, 34 (12), pp.2116-2122. 10.1093/bioinformatics/bty040. hal-01863047

HAL Id: hal-01863047
https://hal-univ-rennes1.archives-ouvertes.fr/hal-01863047
Submitted on 28 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
Databases and ontologies

TOXsIgN: a cross-species repository for toxicogenomic signatures

Thomas A. Darde¹,², Pierre Gaudriault¹, Rémi Beranger¹, Clément Lancien¹, Annaëlle Caillarec-Joly¹, Olivier Sallou², Nathalie Bonvallot¹, Cécile Chevrier¹, Séverine Mazaud-Guittot¹, Bernard Jégou¹, Olivier Collin², Emmanuelle Becker¹, Antoine D. Rolland¹ and Frédéric Chalmel¹,*

¹ Univ Rennes, Inserm, EHESS, 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 sIgNa-tures.

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-
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, i.e., gene expression profiles in response to chemical exposure

The toxicogenomics resources, including CTD and NIH LINCS (Duan, 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, i.e., 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 stand-alone 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
3 Results

3.1 TOXslgN overview and current content

TOXslgN 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 TOXslgN, 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, TOXslgN 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) (Thiése, 2014). TOXslgN 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 TOXslgN repository includes 754 projects for 911 transcriptional 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 TOXslgN, 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 TOXslgN 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.

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 hepatotoxins.

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. 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.

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
We first showed a significant association between the toxicogenomic and toxicological distances ($r^2 = 0.051$, $P < 2.2 \times 10^{-16}$) (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 ($r^2 = 0.056$, $P < 2.2 \times 10^{-16}$; 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, estradiol, mestranol, and β-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).

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 HomoloGene IDs measured, $R$ the total number of HomoloGene IDs meeting the criterion in this signature, $n$ the total number of HomoloGene IDs in the selected signature, $r$ the number of HomoloGene IDs meeting the criterion in this signature, and the Ratio corresponds to $r/[\text{union}(n, R)]$. 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 toxicogenic 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://isatools.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


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


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.


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


