Published in final edited form as:
Toxicology. 2019 January 15; 412: 89–100. doi:10.1016/j.tox.2018.11.005.

Ontology-based Semantic Mapping of Chemical Toxicities

Rong-Lin Wang[a,*], **Stephen Edwards**[b], **Cataia Ives**[b]

[a]Exposure Methods and Measurements Division, National Exposure Research Laboratory, US EPA, Cincinnati, Ohio 45268, USA

[b]Research Computing Division, RTI International, Research Triangle Park, NC 27709, USA

Abstract

This study was undertaken to evaluate the use of ontology-based semantic mapping (OS-Mapping) in chemical toxicity assessment. Nineteen chemical-species phenotypic profiles (CSPPs) were constructed by ontologically annotating the toxicity responses reported in more than seven hundred published studies of ten chemicals on six vertebrate species. The CSPPs were semantically compared to more than 29000 publicly available phenotypic profiles of genes, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, and diseases based on a cross-species phenotype ontology. OS-Mapping was shown to differentiate chemical toxicities among themselves as well as within and across species. It also revealed cases of chemical by species interactions. In addition to confirming similar MOAs (mechanisms of action) for a few chemicals, OS-Mapping also generated novel insights into the MOAs underlying some seemingly different, yet phenotypically similar, classes of chemicals. The nature of a unified cross-species phenotype ontology and its representation of diverse knowledge domains allowed the construction of a complete phenotypic continuum for the 17α-ethynylestradiol_fathead minnow across the biological levels of organization, which complemented a similar one derived from the Comparative Toxicogenomics Database but based primarily on 17α-ethynylestradiol-induced molecular phenotypes. Overall, OS-Mapping has been demonstrated to offer a powerful approach to help bridge the gap between the molecular and non-molecular phenotypes of chemicals characterized by using high throughput or traditional omics methods and their apical endpoints of greater regulatory relevance, which are typically phenotypes found at the higher levels of biological organization. OS-Mapping also enables comparative toxicity assessment among chemicals, both within and across species. Furthermore, the semantic analysis of phenotypes can reveal additional novel MOAs for some well-known chemicals and discover candidate MOAs for chemicals that are less molecularly characterized. A full phenotypic continuum based on OS-Mapping will also be conducive to the future development of adverse outcome pathways. As phenomics continues to advance and the ontological annotation of literature becomes more automated, the power of OS-Mapping will be further enhanced.

[*]Corresponding author: Rong-Lin Wang, 26 W Martin Luther King Dr, Cincinnati, OH 45268, USA, wang.rong-lin@epa.gov, 513-569-7862.

Conflict of interest
The authors declare that there are no conflicts of interest.

Keywords

chemical toxicity; ontology; phenotype; phenomics; semantic analysis

1. Introduction

A shift toward a new paradigm in toxicology is well underway, one that is moving away from animal testing and toward an approach with a greater focus on *in vitro*, short term *in vivo*, and *in silico* tests; high-throughput screening; and toxicity pathways (NRC, 2007). Contemporary advances in omics, systems biology, robotics, computational chemistry, and bioinformatics have been instrumental in the evident progress that has been made under this new paradigm. Programs such as the US Tox21 and EPA's ToxCast (Merrick et al., 2015; Richard et al., 2016) have screened thousands of chemicals for biological effects using biochemical- or cell-based assays and identified numerous molecular targets of potential interest. Transcriptomics, proteomics, and metabolomics-based studies of chemical exposures in both laboratory and field settings have also proliferated, resulting in the determination of many impacted genes, proteins, metabolic pathways, other non-molecular phenotypes, and the subsequent discovery of some novel molecular mechanisms of action (MOAs) for these chemicals and the development of their biomarkers. From these efforts, many powerful bioinformatic approaches for mining high-dimensional omics data in large volumes have also emerged.

Several daunting challenges remain during this paradigm shift. First, high-throughput screenings and omics studies of chemicals focus primarily on molecular phenotypes. Due to the enormous complexity of the phenotypic space and the many-to-many relationships commonly found in genotype-phenotype mapping (Houle et al., 2010), it is difficult to connect molecular phenotypes with their higher level apical endpoints such as anomalies in development, mating behavior, organ size/histopathology, reproduction, and mortality. Second, although some of these non-molecular phenotypes are characterized and recorded as descriptive texts in such studies, they are not readily computable for integrated analysis with molecular data. Third, toxicological studies are most likely conducted on a few model species; it is generally difficult to map and integrate phenotypes from a test species to other species of greater environmental and health concerns. Clearly, to realize the full potential of the new toxicology paradigm, a strategy to address these barrier issues needs to be developed.

Phenomics is the systematic study of the phenotypic responses of an organism on a genome-wide scale, which are determined by the complex interactions of its genotypes and environmental conditions. Phenomics adds a new dimension to the discipline of systems biology, which is largely founded on traditional omics. The term is sometimes used interchangeably with high-throughput phenotyping. Among the most high-profile phenomics projects to date is perhaps the one led by the International Mouse Phenotyping Consortium (Brown et al., 2018), in which mouse genes are systematically mutated and phenotyped by using a wide range of technologies. These phenotypes are described as free text and are then made computable by annotating them with the appropriate ontology terms. With numerous

domain ontologies developed in recent years, phenomics coupled with ontology-based knowledge representation is increasingly being adopted by the scientific community as an approach to dissect complex traits, to better understand genetic variations, and to integrate phenotypes across biological levels of organization and species boundaries (Washington et al., 2009; Houle et al., 2010; Mungall et al., 2010; McMurry et al., 2016; Brown et al., 2018). As such, it should also help to fill the gap between the molecular and non-molecular phenotypes generated from traditional omics platforms and the apical endpoints of greater regulatory significance. The numerous phenomics projects underway are responsible for the abundant and increasing amount of phenotypic information now available across a wide range of species and knowledge domains (Smith et al., 2007; http://www.obofoundry.org; http://www.informatics.jax.org).

The emergence and widespread application of biological ontologies over the last two decades largely coincided with the omics revolution because of the growing necessity to conceptualize, represent, share, and compute this enormous amount of biological knowledge, including phenotypes, across domains and species (Gruber 1995; Ashburner et al., 2000). An ontology represents a knowledge domain, such as those of omics, anatomies, behavior, disease etc., by a set of predefined web ontology language (OWL) constructs and standard vocabulary (Bard and Rhee, 2004). Such a representation standardizes knowledge and makes it accessible to both human and machine inferencing. In addition to pre-composed ontology classes, that is, sets of objects with common attributes originally created by ontology developers, complex phenotypes from phenomics projects can be post-composed into custom ontology classes during annotation by selecting terms from reference domain ontologies using Entity-Quality (EQ) syntax (Washington et al., 2009; Mungall et al., 2010; Hoehndorf et al., 2011; Gkoutos et al., 2017).

Ontologies covering a wide variety of biological domains have many applications, including genome annotation, interpretation of omics findings, knowledge integration across species and biological levels of organization, information retrieval, and semantic computing (Bard and Rhee, 2004; Hoehndorf et al., 2015). An ontology is intrinsically a directed acyclic graph, one in which nodes represent ontology classes and edges denote their subsuming relationships. The information content (IC) of a node is determined by its relative position in the graph: a node with more parent nodes and fewer leaf nodes (i.e., farther away from the root node) has higher IC, thus being more specific and informative. Two nodes are semantically more similar to each other when they share more information, as reflected by the greater IC of their most informative common ancestor (MICA). Subgraphs of an ontology, each representing the phenotypic profile of a chemical, a gene, a pathway, a disease, or other entities of interest by a group of ontology nodes, can be compared to one another for semantic similarities based on various arithmetic manipulations of their underlying ICs. An implicit assumption of this approach of ontology-based semantic mapping (OS-Mapping; Washington et al., 2009) is that, when two such subgraphs are similar, their associated chemicals, genes, pathways, or diseases must share convergent biological mechanisms (Gkoutos et al., 2017).

Although originated from and largely driven by applications in biomedical sciences (McMurry et al., 2016), OS-Mapping could potentially facilitate chemical toxicity

assessment as well. To date, one of the most pressing issues in this area continues to be the need to efficiently evaluate the toxicities of numerous chemicals. Bridging the gap between molecular and non-molecular phenotypes from high-throughput screening/omics studies and the apical endpoints of relevance should facilitate that effort. Comparative assessments of multiple chemicals of their toxicity responses both within and across species will further advance model species-based toxicology as well. In many ways, OS-Mapping appears to represent an ideal approach to implement these types of integrated analysis of toxicological data, as a unified, multi-domain, multi-species ontology can be formed by merging a wide range of public bio-ontologies relevant to toxicology. These include ontologies of gene, protein, chemical, cell, disease, behavior, phenotypes, and importantly, both species-specific and universal anatomies (http://obofoundry.org). After subsuming relationships are established among various classes in the unified ontology by using a reasoner, a type of software tool for inferencing, a semantic analysis can be conducted among individual nodes, or subgraphs of multiple nodes associated with the biological entities of interest. The nature of this unified multi-domain, multi-species ontology allows chemical toxicities and their biology to be dissected both within and across species.

The goals of this pilot study were to evaluate OS-Mapping for its potential applications in chemical toxicity assessment and to demonstrate its value as a computational approach to complement current omics technologies widely used in studying chemical exposures. If proven successful, OS-Mapping would allow a vast amount of non-molecular phenotypes from phenomics projects across several species to inform chemical toxicity assessment in the future. The specific aims of this study were to: 1) build a number of custom chemical-species phenotypic profiles (CSPPs), each of which would summarize multiple published exposure studies for a specific chemical and vertebrate species using post-composed ontology classes; 2) assemble a comprehensive collection of publicly available phenotypic profiles (hereafter referred to as “profiles”) for human, mouse, and zebrafish, with each profile associated with a gene, a biological pathway, or a disease; 3) develop a Java application for the semantic analysis of ontology classes and profiles via OS-Mapping; and 4) assess the performance of OS-Mapping in dissecting chemical toxicities by comparing CSPPs against themselves, and against the assembled public target profiles. During semantic analysis, each CSPP or profile containing multiple ontology classes is effectively an equivalent of an ontology subgraph.

2. Material and Methods

2.1. Ontological annotation of literature and preparation of CSPPs

The US EPA’s ECOTOXicology Knowledgebase (ECOTOX; https://cfpub.epa.gov/ecotox/) and its predecessors were created in the early 1980s to collect, compile, and annotate literature on single chemical exposure studies on a variety of ecological species, including both aquatic and terrestrial animals and plants. The manual annotations of literature by ECOTOX curators were made with predefined codes covering various aspects of an exposure study, including the life stages of experimental organisms, effect types (genetics, morphology, growth, etc.), sites and trends of responses, measurements taken, and the

statistical significance of findings (S. File 1, S. File 2). As of September 2016, ECOTOX had covered over 1600 vertebrate taxa involving more than 7000 chemicals.

The first step in the preparation of CSPPs was selection of the chemicals and species. To ensure that each CSPP would contain a sizeable number of ontology classes, the chemicals and species that ranked high in the number of exposure test results in ECOTOX (S. File 1) were reviewed and selections were made from those remaining after filtering out statistically insignificant and/or phenotypically uninformative mortality-related results. Where possible, preferences were also given to those chemicals studied in more than one species to facilitate the interspecific comparison of toxicity responses of the same chemicals. Ultimately, ten chemicals that had been tested in six vertebrate species that met these criteria were selected for ontological annotations: atrazine, bisphenol A, cadmium chloride, chlorpyrifos, copper sulfate, cypermethrin, dioxin, 17α-ethynylestradiol, malathion, and Tris(1,3-dichloroisopropyl) phosphate (TDCPP; Table 1). The six vertebrate species were: carp (*Cyprinus carpio*), zebrafish (*Danio rerio*), fathead minnow (*Pimephales promelas*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), and trout (*Onchorhynchus mykiss*).

The next step of CSPP preparation was to export and preprocess the annotations of selected combinations of chemical and species from ECOTOX (Figure 1A). A text file was exported for each chemical and species containing information about the response site, measurements, response trend, effect types, statistical significance of test results, organism life stages, and the references of original source publications (S. File 2). Statistically insignificant and/or mortality-related results were filtered out. A few publications with obvious quality issues or without annotatable phenotypes were excluded. After these steps, a non-redundant set of 726 publications spanning the last several decades was identified for subsequent manual annotations using EQ syntax (Figure 1B, 1C). These publications were deemed acceptable because they originated largely from peer-reviewed journals and had been previously screened by ECOTOX curators. The original annotations in ECOTOX by its predefined codes were retained for reference purposes only. These annotations were not used directly in the EQ statements because of several considerations: a wide range of granularity in the coded responses, frequent omissions of histological details and other findings reported in the original publications, and occasional errors. These issues existed primarily because ECOTOX was not originally designed for ontological annotations and because ECOTOX had evolved over almost four decades under the likely curation of many different individuals over its annotation history. Nevertheless, these issues made it difficult to directly incorporate ECOTOX annotation codes into EQ statements with high specificity and accuracy.

In its most basic form, EQ syntax describes how an entity (E) such as an anatomical part, a biological process, or a biological function, is altered in its quality (Q; Figure 1C; Washington et al., 2009; Hoehndorf et al., 2011; Gkoutos et al., 2017). The ontology term for an entity is selected from reference domain ontologies such as the Gene Ontology, Cell Ontology, and various anatomy ontologies. The quality term comes from the Phenotype And Trait Ontology (PATO; http://purl.obolibrary.org/obo/pato.owl). An entity and a quality are related to each other by an object property (R) from the Relations Ontology (RO; http://purl.obolibrary.org/obo/ro.owl). At times, EQ syntax needs to be expanded to include a secondary entity, as in E_1-Q-E_2, where Q denotes a relational quality. For example,

"blood_serum(E_1) has_quality(R) increased_ concentration (Q) towards (R) glucose (E_2)" denotes phenotype "raised blood glucose level". For a more complex phenotype, an entity can also be further post-composed, giving rise to a more generalized syntax in the form of [E_{1a}-R-E_{1b}]-[Q-QL]-[E_{2a}-R-E_{2b}], where QL is a quality modifier (e.g., PATO_0000460, "abnormal"). The post-composition of an entity may extend over multiple levels, which may either be nested or parallel to one another. To provide a more accurate annotation, an entity term could also be modified by another quality term, as in [E-R-Q]-[Q].

The chemical toxicity responses in the 726 publications selected were manually curated and annotated with EQ syntax by using Phenote software (Phenote_1_8_13_windows-x64_install4j.exe, released 11-29-2012; http://www.berkeleybop.org/index.html). Phenote was custom configured, and all user-designated ontologies in Open Biomedical Ontology (OBO) format were preloaded into memory during each session to expedite the selection of specific and appropriate ontology terms for a given phenotype. Once the annotation work was complete for all the selected chemicals/species, their tab-delimited outputs were manually checked to ensure that the order of terms in a post-composed entity remained unchanged (S. File 3). This step was necessary because Phenote occasionally rearranged those terms upon saving and reopening an annotation file. EQ annotations were then converted into ontology classes in Manchester syntax (https://www.w3.org/TR/owl2-manchester-syntax) by using a custom Perl script (Figure 1D, S. File 4), and further into RDF/XML format (https://www.w3.org/TR/rdf-syntax-grammar/) by using Protégé (http://protege.stanford.edu). Each result reported in a publication was in effect annotated into a custom ontology class with a distinct ID. All ontology classes constructed for each chemical and species were syntactically unique. A CSPP was then prepared consisting of all the relevant class IDs associated with a given combination of chemical and species (Figure 1E). A total of 19 CSPPs was created from annotating the 726 publications.

2.2. Assembly of profiles from public phenomics data

Profiles were prepared for human, mouse, and zebrafish, the three vertebrate species for which public omics data, including phenotypes, is likely the most abundant. All data were downloaded on September 15, 2017. For the human profiles, two files were acquired from the Human Phenotype Ontology (HP) site (http://human-phenotype-ontology.github.io/downloads.html, redirected to http://compbio.charite.de/jenkins/job/hpo.annotations.monthly/lastStableBuild).

The first file, "ALL_SOURCES_ALL_FREQUENCIES_genes_to_phenotype.txt", contained human genes, each of which was annotated with multiple HP terms. A second file, "ALL_SOURCES_ALL_FREQUENCIES_diseases_to_genes_to_phenotypes.txt", linked human diseases to both human genes and HP terms. The human diseases were coded with either OMIM (Online Mendelian Inheritance in Man) IDs or Orphanet IDs (http://www.orpha.net/consor/cgi-bin/index.php?lng=EN). The mouse data file, "MGI_GenePheno.rpt", was downloaded from the Jackson Laboratory (http://www.informatics.jax.org/downloads/reports/index.html#pheno). Each mouse gene in the file was linked to multiple Mammalian Phenotype Ontology (MP) terms. The second mouse data file, "MGI_Geno_DiseaseDO.rpt", linked the OMIM disease IDs to MP terms. For the

zebrafish profiles, genes linked to Zebrafish Phenotype Ontology (ZP) terms were downloaded (https://github.com/Phenomics/zebrafish-phenotype-ontology-build/blob/master/annot_gene_pos.txt). In addition to the profiles anchored by genes and diseases, profiles were also prepared for mouse KEGG (Kyoto Encyclopedia of Genes and Genomes; http://www.genome.jp/kegg) pathways by replacing the gene members in each pathway with their associated MP terms. A total of 29154 profiles was assembled, including 3498 by human genes in HP terms, 12009 by mouse genes in MP terms, 5892 by zebrafish genes in ZP terms, 1987 by OMIMs in MP terms (denoted as M_OMIM_digits), 4164 by human diseases and disorders in HP terms (denoted as OMIM_digits), 1272 by rare diseases in HP terms (Orpha_digits), 313 by KEGGs in MP terms, and 19 CSPPs.

2.3. Development of an OS-Mapping Java application

After evaluating a few existing applications available for ontology-based semantic analysis (e.g., OwlSim, http://www.berkeleybop.org/software/owlsim; SML Toolkit, http://www.semantic-measures-library.org/sml/index.php?q=toolkit; GOSemSim, Yu et al., 2010), a more streamlined and flexible application, called OS-Mapping.java, was developed in-house to better meet our needs in analyzing chemical toxicities. As a command line tool, OS-Mapping.java relies on several Java packages including OWLAPI (version 4.2.5; Horridge and Bechhofer, 2011), Semantic Measure Library (SML, version 0.9.4d; Harispe et al., 2014), and various reasoners (Dentler et al., 2011; Kazakov et al., 2011; Mendez, 2012; Ceylan et al., 2015). The OWLAPI is a well-known and widely adopted application programming interface that provides various functionalities to the creation and manipulation of OWL ontologies. The SML works with a preexisting reasoned ontology and provides multiple semantic measures for both pairwise and group-wise analyses of ontology classes. A reasoner examines ontology axioms (statements about classes, individuals, or properties) for their logical validity (consistency and satisfiability), and then infers subsumption relationships among classes. The development of OS-Mapping.java was conducted within NetBeans, an integrated development environment (version 8.2; https://netbeans.org).

To evaluate the statistical significance of the semantic similarity scores between a query and its target profiles, OS-Mapping.java generates a user-specified number of random profiles of the same size as the original query and calculates their similarities against target profiles individually. A random profile is formed by sampling all the classes present in a merged ontology, which, in the current version of cross-species phenotype ontology (http://purl.obolibrary.org/obo/upheno/vertebrate.owl; Köhler et al., 2013), number more than 151000. In effect, a distribution of semantic similarity scores is generated for this query against each of the respective target profiles using the same group of random profiles. The scores at the top 5% and top 1% maximum range in each ascendingly sorted distribution are then selected as the cutoffs for the corresponding pair of query and target profiles. For the next query profile, a new set of random profiles is generated based on its size, which is then compared against all the target profiles again. The process repeats until similarity score cutoffs have been generated for all possible comparisons of query-target profiles. A pair of query and target profiles would be declared significantly similar to each other if its similarity was greater than its associated cutoffs. Although informative, this procedure is computationally expensive, particularly when a query has a large size and the number of

random profiles grows. In addition, there is a possibility, depending on the query size and number of random profiles generated, that some sampled classes in a random profile may be correlated to various extents due to the nature of an ontology graph. Such a correlation, when it occurs, would violate one of the assumptions of resampling that all individuals in a population should have an equal probability to be sampled.

When conducting an analysis, OS-Mapping.java takes a single configuration file (S. File 5) at the command line as the input argument to specify such analysis parameters as a starting root ontology, a set of query and target profiles, and the choices among seven reasoners, ten information content measures, 22 pairwise, and 13 group-wise semantic similarity measures. A preexisting reasoned ontology could be provided in lieu of the root ontology. Each profile contains four required fields, in the following order: a profile ID, a profile definition, an ontology class ID, and an ontology class definition (Figure 1E). Additional fields in a profile are ignored. A profile could be anchored by a gene, disease, biological pathway, chemical, or any other biological entities of interest. Common reasoners supported include Hermit, Pellet, Elk, Jfact, Snorocket, Jcel, and Born (Dentler et al., 2011; Kazakov et al., 2011; Mendez, 2012; Ceylan et al., 2015).

2.4. Semantic similarity measures

The following four measures implemented in SML were selected for this study based on the previous evaluation of their performance on comparing gene products (Pesquita et al., 2009): 1) information content, “ICI_SANCHEZ_2011” (Sanchez et al., 2011); 2) pairwise similarity between two OWL classes, “SIM_PAIRWISE_DAG_NODE_LIN_1998” (Lin, 1998); 3) direct group-wise measure, “SIM_GROUPWISE_DAG_GIC” (Pesquita et al., 2007); and 4) indirect group-wise measure, “SIM_GROUPWISE_BMA” (Pesquita et al., 2008). They are defined as follows:

1. “ICI_SANCHEZ_2011”

$$IC(\mu) = -\log\left(\frac{\frac{|leaves(\mu)|}{|A(\mu)|}+1}{max\_leaves+1}\right)$$

Where $IC(\mu)$ denotes information content for class node μ; $leaves(\mu)$, the number of leaf nodes below μ; A (μ), the number of parental nodes above μ; and max_leaves, the number of leaf nodes below the root. A class with more parental nodes and fewer leaf nodes has a greater IC.

2. “SIM_PAIRWISE_DAG_NODE_LIN_1998”

$$sim(\mu, v) = \frac{2 \times IC\left(MICA_{\mu, v}\right)}{IC(\mu) + IC(v)}$$

Where $MICA_{\mu,v}$ denotes the most informative common ancestor of the two nodes. The numerator measures the commonality between the two nodes, while the denominator measures their respective distances from *MICA*.

3. "SIM_GROUPWISE_DAG_GIC"

$$simGIC(A,B) = \frac{\sum_{i=1}^{N} IC(t_i)}{\sum_{j=1}^{M} IC(t_j)}$$

Where $t_{i=1\text{-}N}$ and $t_{j=1\text{-}M}$ are the terms in the intersection and union of group A and B.

4. "SIM_GROUPWISE_BMA"

$$simBMA(A,B) = \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N} maxS_{ij} + \frac{1}{M}\sum_{j=1}^{M} maxS_{ij}\right),\ 1 \le j \le M; 1 \le i \le N$$

Where groups A and B have size N and M, *respectively*, and S_{ij} is a pairwise similarity matrix with N rows and M columns for all possible class pairs between A and B based on a semantic measure, in this case, "SIM_PAIRWISE_DAG_NODE_LIN_1998". *simBMA* represents the average of the best row scores (A compared to B) and best column scores (B compared to A) in the similarity matrix for groups A and B.

2.5. OS-Mapping analysis

A merged and reasoned ontology was prepared during the semantic analysis of the 19 CSPPs against a total of 29154 target profiles (Figure 2A). The beginning root ontology was set as the vertebrate.owl (as of February 5, 2018), a cross-species phenotype ontology encompassing HP, MP, and ZP. It, in turn, imported many other domain ontologies in anatomy, behavior, cell, chemical, disease, gene, pathology, phenotypes, protein, relations etc. These domain ontologies had different release dates; for example, the imported ro.owl was released on April 12, 2015 while the hp.owl was released on January 26, 2018. Also imported were the custom ontology classes constructed from the EQ annotations of toxicity phenotypes ("19 chemical-species.owl") and the external reference ontology terms contained therein ("OS-Mapping.import.owl"). The latter was generated online by using OntoFox (Xiang et al., 2010) with appropriate terms from individual reference ontologies as top-level sources and additional import options of "includeAllIntermediates" and "includeAllAxiomsRecursively". Once all the ontologies were imported and merged into a single ontology, it was reasoned with Elk reasoner, a widely recognized high performer on large ontologies (Kazakov et al., 2012). The resultant ontology graph (sampled in Figure 2B, 2C), containing over 151000 classes and 1.7 million logical axioms, enabled the SML Engine to conduct semantic analysis. In this graph, multiple classes contained in a CSPP query are represented by a set of nodes (Figure 2D, solid circles), while target profiles are various combinations of selected nodes in different sizes. In effect, when OS-Mapping compares a CSPP to target profiles, semantic similarities are calculated for groups of nodes of interest pairwise, as shown in a simplified sample output (Figure 2E).

To determine if a pair of query and target profiles was statistically similar to each other, the parameter “GENERATE_SIM_SCORE_DISTRIBUTION” was turned on along with “NUM_SIMULATED_GRPS” set at 500 (S. File 5). In other words, for each such pair, $P_{0.05}$ and $P_{0.01}$ cutoffs (i.e., P-value ≤ 0.05, ≤ 0.01) were determined based on its unique distribution of 500 ascendingly sorted similarity values, which were calculated from 500 random profiles against the target profile under analysis. Those query-target profiles with similarities greater than their respective $P_{0.05}$ cutoffs were retained for further consideration. All analyses were conducted on the US EPA ATMOS Linux cluster (Intel E5–2697A 2.6 GHz processors), taking a combined total of 3158 CPU hours.

Several follow-up analyses were also conducted. The scope of coverage of mammalian phenotypes by the CSPPs was estimated as follows. First, the MP terms having pairwise similarities to the custom ontology classes in each CSPP ≥ 0.7, 0.8, or 0.9 were identified. Each set of MP terms was then treated as seeds to extract their superclasses (i.e., parental classes) in mp.owl using the Robot tool (http://robot.obolibrary.org). The phenotypic coverage of each CSPP was measured by the number of the 27 categories of high level MP phenotypes (http://www.informatics.jax.org) present in its associated superclasses. To determine the relationships of the CSPPs, their similarities were evaluated in the software Cytoscape (http://cytoscape.org). A complete phenotypic continuum across biological levels of organization was constructed for the 17α-ethynylestradiol_fathead minnow by organizing its top ten matched chemicals, genes, KEGG pathways, and diseases. Additional best matched phenotypes at the levels of biological processes, cell, tissue/organ, and organism were also identified from the five highest scoring HP/MP/ZP terms based on the pairwise similarity between the individual classes in the 17α-ethynylestradiol_fathead minnow and all the classes present in the target profiles. For comparison, a similar phenotypic continuum was also constructed for 17α-ethynylestradiol based on the information available as of July 2018 in the Comparative Toxicogenomics Database (CTD; http://ctdbase.org/).

3. Results

In this study, we assembled a total of 19 CSPPs containing nearly 4000 ontology classes that were manually annotated and constructed from chemical toxicity responses (Figure 1). These CSPPs were semantically compared to each other to identify chemicals that showed similar toxicities, as well as compared against a broad array of over 29000 target profiles corresponding to genes, KEGG pathways, and diseases to highlight potential MOAs (Figure 2A, S. File 6). To evaluate the performance of OS-Mapping as a phenotype-oriented approach to dissect chemical toxicities, we will first focus on the relationships revealed by the CSPPs in general. We will then examine the CSPP corresponding to 17α-ethynylestradiol phenotypes observed in the fathead minnow in more detail to evaluate the mechanistic information provided by OS-Mapping. Unless specified otherwise, a statement about genes, pathways, chemicals, diseases, and their mutual similarities refers to their respective profiles.

The CSPPs appeared to have a good coverage of the 27 categories of high level mammalian phenotypes (Figure 3). As expected, the lower threshold of minimum similarities between the MP terms and the custom ontology classes in each CSPP resulted in a greater coverage

of the phenotypes, as the effect of retaining more MP terms initially was the extraction of more superclasses/phenotypic categories. The coverages were 95%, 85%, and 77%, respectively, when the minimum similarities were set at 0.7, 0.8, and 0.9. The coverage reached 100% when the minimum similarity was lowered to 0.5. The evaluation of individual MP terms mapped to the custom ontology classes from the CSPPs at various similarities suggested that the minimum threshold of 0.7 was sufficiently high for assessing the mammalian phenotypic categories covered by the CSPPs. It is notable that the category of mortality/aging was the least covered here because mortality responses were deemed uninformative and omitted during the ontological annotations of the published studies. The rat-based CSPPs tended to have a better coverage overall. The CSPPs representing the fewest phenotypic categories were cypermethrin_carp (21) and TDCPP_zebrafish (22).

The CSPPs mapped to a wide range of target profiles, with their specificities varying by more than 100-fold as measured by the number of hits (Table 1). Most of the CSPPs resembled one another to some degrees, as indicated by the number of other similar CSPPs mapped. A majority (3702/4462 = 83%) of the CSPP-mapped genes was from mouse, with the remainder split between human and zebrafish genes. Ultimately, the CSPPs were mapped to approximately 21% (4462/21399) of the genes under study. By species, these percentages were 31% (3702/12009) for mouse, 12% (422/3498) for human, and 6% (338/5892) for zebrafish, respectively. The CSPPs also matched to approximately 16% (1181/7423) of the diseases. Individually, cypermethrin_rat, malathion_mouse, and atrazine_mouse had the highest number of hits, whereas TDCPP_zebrafish, atrazine_zebrafish, copper sulfate_fathead minnow, and bisphenol A_zebrafish had the fewest hits. This trend was observed across species in genes, pathways, and diseases. The size of a CSPP did not appear to be correlated with its number of mapped targets ($r = 0.41$).

The relationships among CSPPs were complex, as shown in their network with individual CSPPs as nodes and edges weighted by their indirect group-wise similarity scores (value range 0.0 to 1.0; Figure 4). The same chemicals in different species, for example, malathion_mouse vs malathion_rat (0.74), atrazine_rat vs atrazine_mouse (0.71), and 17α-ethynylestradiol_fathead minnow vs 17α-ethynylestradiol_zebrafish (0.69), were highly similar, as expected. In addition, different chemicals, such as chlorpyrifos_carp vs atrazine_carp (0.77), atrazine_rat vs malathion_rat (0.74), and malathion_rat vs cypermethrin_rat (0.72), mapped to each other very well too. In fact, chlorpyrifos_carp vs atrazine_carp achieved the highest similarity score among the CSPPs other than those self-hits. Overall, these CSPPs tended to segregate into two distinct groups. Atrazine_zebrafish, copper sulfate_fathead minnow, bisphenol A_zebrafish, and TDCPP_zebrafish belonged to the group with relatively low similarities to others, and the remainder of the CSPPs appeared to share much greater similarities among themselves. Zebrafish appeared to be less sensitive to the chemicals relative to the other species except when evaluating estrogenic compounds (Table 1, Figure 4). Interestingly, atrazine and copper sulfate were present in both groups, a strong indication of chemical by species interaction in their toxicity responses. The lowest yet still statistically significant similarity was found between malathion_mouse and copper sulfate_fathead minnow, at 0.46.

Given the complexity of chemical-induced toxicity responses, it would be instructive to organize them according to their intrinsic biological hierarchy. As a well-studied semisynthetic estrogen and endocrine disrupting chemical, 17α-ethynylestradiol provided a good sample for this exercise (Figure 5). This CSPP contained 156 custom-constructed ontology classes covering phenotypes across multiple biological levels from 34 studies. As expected, its semantically mapped targets were also diverse. All the top ten genes were involved, to various extents, in some aspects of reproductive biology: gonad development, male and female meiosis, endocrine glands, sperm and oocyte production, or infertility, as demonstrated by their mutant/allelic analyses (repro15 and repro16, JAX 2004; Ccnb1ip1, Dms, Cdk16, Cnot7, Stra8, Immp2l, Tmem203, Ccdc42, http://www.informatics.jax.org). In fact, out of the 915 genes mapped to the 17α-ethynylestradiol_fathead minnow, 788 were mouse genes linked to high level MP terms, and 89% of them (703/788) were involved in reproductive systems (MP_0005389; http://www.informatics.jax.org). In contrast, across the mouse genome overall, this value was only 21% (2411/11409). At the pathway level, although none of the top ten pathways were significant ($P_{0.05}$), all were involved in reproductive processes and functions as well, according to mutant/allelic analyses of their gene members (http://www.informatics.jax.org). The sole significant pathway, mmu00440 (Phosphonate and phosphinate metabolism), was ranked low by similarity scores (263th out of 313), but three of its six gene members were also associated with reproductive system phenotypes. Moreover, the top phenotypes mapped by the 17α-ethynylestradiol_fathead minnow were also examined at the levels of biological process, cell, tissue/organ, and organism by identifying the HP, MP, and ZP terms of the target profiles most semantically similar to the individual terms of the 17α-ethynylestradiol_fathead minnow. Not surprisingly, almost all those terms related to reproduction overlapped with the profiles of genes, pathways, and diseases linked to the 17α-ethynylestradiol_fathead minnow. Among the notable biological processes mapped to the 17α-ethynylestradiol_fathead minnow were female and male meiosis, spermatogenesis, oocyte maturation, and fertilization. Regarding disease, most of the top matches were reproductive disorders. As to the relationships to the other CSPPs, in addition to the expected mapping to the 17α-ethynylestradiol _zebrafish (0.69), the 17α-ethynylestradiol_fathead minnow also shared substantial similarities to several chemicals with seemingly different MOAs, such as atrazine_rat (0.65) and copper sulfate_rat (0.63).

A phenotypic continuum for 17α-ethynylestradiol was also constructed by extracting the data curated and compiled by the CTD (Figure 6). The CTD mappings of 17α-ethynylestradiol to GO processes, pathways, most of the diseases, and other chemicals were derived from the 17α-ethynylestradiol-gene interactions curated from the literature. The 17α-ethynylestradiol to phenotypes, as delineated by biological process, cell, tissue/organ, and organism, and to the remaining diseases came directly from the literature curation. Almost all 17α-ethynylestradiol-gene interactions (98%) in the CTD were based on the chemical-induced changes in gene expressions. Overall, many 17α-ethynylestradiol phenotypes from the CTD at the levels of cell, tissue/organ, and organism were related to immune functions and reproductive biology. This trend was somewhat aligned with the top matched 17α-ethynylestradiol genes (e.g., ESR1, ESR2, PGR) and chemicals (e.g.,

bisphenol A, estradiol), but less clear in the top GO processes, pathways, diseases, and phenotypes at the process level.

4. Discussion

OS-Mapping represents a unique approach to the study of chemical toxicities. It complements a typical omics study by providing a new perspective of semantically similar phenotypes at multiple levels of biological organization. It can also provide information about the underlying genotypes and molecular mechanisms of a biological target previously characterized only phenotypically. This pilot study explored the toxicological applications of OS-Mapping by converting previously published toxicity responses to chemicals into CSPPs of multiple ontology classes, and then comparing them semantically to the numerous profiles of genes, pathways, and diseases, and to one another, leading to a better understanding of the underlying MOAs.

OS-Mapping can differentiate chemical toxicities both within and across species. Given the 95% average coverage rate of the 27 categories of high level mammalian phenotypes (Figure 3), the CSPPs should be quite comparable. The fact that CSPPs mapped to more mouse genes than human and zebrafish genes is likely just a reflection of the mouse phenome that is better characterized. The CSPPs appear to contain both shared and specific toxicity responses. The shared responses are indicated by the fact that most of these CSPPs were similar to one another to some degree (Table 1, Figure 4). The specificity of CSPPs is evident in the wide variation in the number of genes, pathways, and diseases mapped by each CSPP. Conceivably, the toxicity responses of a chemical are determined by both its MOAs and target species. Perturbations at the higher levels of biological organization and in later life stages, on the other hand, may have a homogenizing effect on chemical toxicities, making some CSPPs more similar to one another. As to the degree of toxicity, it may be reasonable to assume that the more targets a chemical maps to, the more toxic it is. Under this assumption, therefore, 17α-ethynylestradiol would be considered more toxic to fish (i.e., wider impact) than copper sulfate, atrazine, or TDCPP; malathion would be more toxic to mouse than chlorpyrifos. In the meantime, there appear to be interactions between chemicals and species as well. Atrazine, thus, appears to be more toxic to rodents than it is to fish. The same conclusion could also be made for cypermethrin and copper sulfate. Chlorpyrifos, on the other hand, may be more toxic to fish than it is to rodents. And, the difference in toxicity to cadmium chloride between fish and rodents appears to be minimal. Overall, these findings are largely consistent with the current knowledge of differential species sensitivity to chemicals (Belanger et al., 2017).

OS-Mapping can yield novel insights into chemical MOAs. There are several notable pairs of highly similar CSPPs: atrazine_carp/chlorpyrifos_carp, atrazine_rat/malathion_rat, malathion_rat/malathion_mouse, atrazine_rat/atrazine_mouse, and 17α-ethynylestradiol_fathead minnow/17α-ethynylestradiol_zebrafish (Figure 4). Although it is not surprising to observe similar interspecific toxicity responses for the same chemicals (malathion, atrazine, 17α-ethynylestradiol), the apparent MOAs of atrazine (a photosystem II inhibitor; Shimabukuro and Swanson, 1969) and chlorpyrifos/malathion (acetylcholinesterase inhibitors; Colovic et al., 2013) offer few clues about what underlies

their shared toxicity phenotypes. The top contributing phenotypes to the similarity between atrazine_carp and chlorpyrifos_carp are acetylcholinesterase activity (brain) and a few key indicators of stress and immune physiology (antioxidant activity, alkaline phosphatase activity in head kidney/kidney; erythrocyte quantity, nitric-oxide synthase activity, nitric oxide level, gene and protein expressions of heat shock protein 70 in spleen). The top KEGG pathways mapped to these two CSPPs, although not significant at $P_{0.05}$, are dominated by those related to immune responses (e.g., mmu05340, primary immunodeficiency; mmu05332, graft-versus-host disease; mmu05330, allograft rejection; S. File 6). In rodents, besides immune responses, the top KEGG pathways mapped to atrazine_rat and malathion_rat also include those involved in reproductive systems (e.g., mmu00592, mmu04913, mmu00140, $P_{0.05}$; Figure 5, S. File 6). Therefore, the MOAs underlying atrazine toxicity in animals appear to resemble those of organophosphates, a hypothesis at least partially supported by the observed synergism between atrazine with both malathion and chlorpyrifos in a mixture (Pape-lindstrom and Lydy, 1997).

OS-Mapping enables the construction of a phenotypic continuum for chemical toxicity assessment. Two such continuums (Figure 5, 6) for 17α-ethynylestradiol offer contrasting views of its toxicities from different perspectives: one based on semantically similar phenotypes across biological levels of organization, and the other largely derived from molecular phenotypes alone (genes with expressions impacted by 17α-ethynylestradiol), except for a small number of higher level phenotypes directly curated from relevant literature. In both continuums, higher level phenotypes (cell, tissue/organ, organism) indicate a significant impact of 17α-ethynylestradiol on reproductive biology. Although this pattern is consistently observed throughout the 17α-ethynylestradiol_fathead minnow continuum, it is not as obvious in the CTD-17α-ethynylestradiol continuum among all its top mappings. Overall, the CTD-17α-ethynylestradiol continuum contained many more genes and their derived pathways/diseases than its counterpart. At the gene level, only 29 mapped genes are shared in both continuums. Noticeably absent in the top ten genes of the 17α-ethynylestradiol_fathead minnow continuum are two estrogen receptors (ESR1, ESR2), which, with similarity scores of 0.62 and 0.57, were deemed insignificant at $P_{0.05}$. Given the well-established functions of these receptors in estrogen signaling, their absence appears to indicate that OS-Mapping is less sensitive than the direct omics assays of gene expression for establishing chemical-gene linkages. This interpretation is also supported by the more than ten-fold difference between the two continuums in the number of genes mapped to 17α-ethynylestradiol. At the pathway level, the top ten in the 17α-ethynylestradiol_fathead minnow continuum contain a substantial number of gene members whose mutations/alleles directly led to phenotypes in reproductive systems (i.e., mapped to MP_0005389; http://www.informatics.jax.org). For example, alpha-Linolenic acid metabolism (mmu00592) was involved in gonad development, spermatogenesis, oogenesis, fertilization, and infertility, as indicated by six of its gene members (Acox1, acyl-Coenzyme A oxidase 1; Fads2, fatty acid desaturase 2; Pla2g3, 4a, 6, 10, phospholipase A2 group III, IVa, VI, X; http://www.informatics.jax.org). Nine out of these ten pathways are also found in the CTD-17α-ethynylestradiol continuum, but ranked hundreds below. For diseases, only two of the top ten from the 17α-ethynylestradiol_fathead minnow continuum, OMIM_616067 (gonadal dysgenesis) and OMIM_614842 (hypogonadism), are present in CTD-17α-ethynylestradiol

disease mappings, but again, with very low rankings. Overall, these comparisons suggest that OS-Mapping appears to perform better for mapping toxicities of a chemical at the higher levels of biological organization, whereas the CTD provides a more comprehensive coverage of genes by relying on the data from omics-based gene expression assays directly.

The validity of OS-Mapping and its values for chemical toxicity assessment are strongly supported and demonstrated in this pilot study by the finding that the 17α-ethynylestradiol_fathead minnow continuum is dominated by reproductive phenotypes throughout. This finding is consistent with the established estrogen MOAs (Nilsson et al., 2001; Hess, 2003). Perhaps one of the most exciting prospects is that, due to the development of many reference ontologies over diverse knowledge domains, modeling chemical toxicities by incorporating a vast and growing amount of computable phenomics data throughout biological levels of organization has been made possible both within and across species. To fully realize this potential, however, efficient curation of toxicity responses into computable ontology classes must be attained. Manual curation of toxicity responses by EQ syntax is not only slow, but also subjective at times. In the current study, the curation of more than 700 publications took one individual several months to complete. Moreover, the phenotypes encountered during post composition were sometimes open to interpretation, leading to possible variations in the final annotations in terms of specificity and accuracy. A preferred solution to these issues would be found in the automated curation of free text, which is an area under active research. Considerable progress has been made for various aspects of this process, including mapping free text to ontology terms (e.g., https://www.ebi.ac.uk/spot/zooma), generating ontology classes en masse by adopting design patterns (Osumi-Sutherland et al., 2017), and even the completely automated construction of full EQ statements with binding relations (Cui et al., 2015). More efforts are needed, however, before the performance of an automated curation tool will be as accurate as that of human curators.

5. Conclusions

This study demonstrated that OS-Mapping offers a powerful approach to help bridge the gap between the molecular/non-molecular phenotypes of chemicals characterized by using traditional omics methods and their apical endpoints of greater regulatory relevance. OS-Mapping also enables the comparative toxicity assessment among chemicals, both within and across species. Furthermore, the semantic analysis of phenotypes can reveal additional novel MOAs for some of the well-known chemicals and assist in the discovery of candidate MOAs for chemicals that are less molecularly characterized. A full phenotypic continuum delineating chemical toxicities will also be conducive to the future development of adverse outcome pathways, a framework increasingly adopted under the new toxicology paradigm (Ankley et al., 2010). Continued advances in phenomics and more automation of the ontological annotation of the literature will further enhance the power of OS-Mapping.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank many members of the ontology community for their valuable advice on reasoners, OWLAPI, and OBO ontologies. Special thanks are due to Stephen Erickson for assistance on setting up a local copy of ECOTOX Database, Colleen Elonen for providing numerous journal articles, Yvonne Bradford and James Balhoff for advice on annotating chemical-induced toxicity responses, and Sébastien Harispe for guidance on the Semantic Measure Library. We also thank Nancy Baker and Kellie Fay for critical reviews of this manuscript. The information in this document has been funded wholly (or in part) by the US Environmental Protection Agency. It has been subjected to review by the National Exposure Research Laboratory and approved for publication. Approval does not signify that the contents reflect the views of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

Abbreviations

CSPP	chemical-species phenotypic profile
CTD	Comparative Toxicogenomics Database
EPA	Environmental Protection Agency
EQ	Entity-Quality
ECOTOX	ECOTOXicology Knowledgebase
GO	Gene Ontology
HP	Human Phenotype Ontology
IC	information content
ID	identification
KEGG	Kyoto Encyclopedia of Genes and Genomes
MICA	most informative common ancestor
MOA	mechanism of action
MP	Mammalian Phenotype Ontology
OBO	Open Biological and biomedical Ontology
OMIM	Online Mendelian Inheritance in Man
OWL	Web Ontology Language
OWLAPI	OWL Application Programming Interface
PATO	Phenotype And Trait Ontology
RDF/XML	Resource Description Framework/Extensible Markup Language
RO	Relations Ontology
SML	Semantic Measure Library
TDCPP	Tris(1,3-dichloroisopropyl) phosphate

ZP Zebrafish Phenotype Ontology

References

Ankley GT, et al., 2010 Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem 29(3):730–41. doi: 10.1002/etc.34. [PubMed: 20821501]

Ashburner M, et al., 2000 Gene ontology: tool for the unification of biology. Nat. Genet 25(1):25–29. [PubMed: 10802651]

Bard JBL, Rhee SY, 2004 Ontologies in biology: design, applications and future challenges. Nat. Rev. Genet 5:213–222. [PubMed: 14970823]

Belanger S, et al., 2017 Future needs and recommendations in the development of species sensitivity distributions: Estimating toxicity thresholds for aquatic ecological communities and assessing impacts of chemical exposures. Integr. Environ. Assess. Manag 13(4):664–674. doi: 10.1002/ieam.1841. [PubMed: 27531323]

Brown S, et al., 2018 High-throughput mouse phenomics for characterizing mammalian gene function. Nat. Rev. Genet 19:357–370. doi: 10.1038/s41576-018-0005-2. [PubMed: 29626206]

Ceylan İİ, et al., 2015 The Bayesian Ontology reasoner is BORN! In Dumontier M, et al., (Eds.), Proceedings of the 4th International Workshop on OWL Reasoner Evaluation (ORE 2015), 1387:8–14, 2015 CEUR Workshop Proceedings.

Colovic MB, et al., 2013 “Acetylcholinesterase Inhibitors: Pharmacology and Toxicology”. Curr. Neuropharmacol 11 (3): 315–335. doi:10.2174/1570159X11311030006. [PubMed: 24179466]

Cui H, et al., 2015 Charaparser+EQ: Performance evaluation without gold standard. Proceedings of the Association for Information Science and Technology, 52 (1), 1–10. doi:10.1002/pra2.2015.145052010020.

Dentler K, et al., 2011 Comparison of Reasoners for large Ontologies in the OWL 2 EL Profile. Semantic Web 2 (2) 71–87. doi: 10.3233/SW-2011-0034.

Gkoutos GV, et al., 2017 The anatomy of phenotype ontologies: principles, properties and applications. Brief. Bioinform doi:10.1093/bib/bbx035.

Gruber TR, 1995 Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum. Comput. Stud 43(5–6), 907–928. doi:10.1006/ijhc.1995.1081.

Harispe S, 2014 The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. Bioinformatics 30 (5), 740–742, doi:10.1093/bioinformatics/btt581. [PubMed: 24108186]

Hess RA, 2003 Estrogen in the adult male reproductive tract: A review. Reprod. Biol.Endocrinol 1:52. [PubMed: 12904263]

Hoehndorf R, et al., 2011 PhenomeNET: a whole-phenome approach to disease gene discovery. Nucleic Acids Res 39 (18) e119, doi:10.1093/nar/gkr538. [PubMed: 21737429]

Hoehndorf R, et al., 2015 The role of ontologies in biological and biomedical research: a functional perspective. Brief. Bioinform 16(6), 1069–1080. doi: 10.1093/bib/bbv011. [PubMed: 25863278]

Horridge M, Bechhofer S, 2011 The OWL API: A Java API for OWL Ontologies. Semantic Web Journal 2(1), 11–21. doi: 10.3233/SW-2011-0025.

Houle D, et al., 2010 Phenomics: the next challenge. Nat. Rev. Genet 11:855–866. [PubMed: 21085204]

JAX Reproductive Mutagenesis Program, 2004 Heritable mouse mutants from The Jackson Laboratory Reproductive Genomics Mutagenesis Program. MGI Direct Data Submission, J:92463. URL: http://reprogenomics.jax.org.

Kazakov Y, et al., 2011 Concurrent Classification of EL Ontologies. Technical report. University of Oxford 2011.

Kazakov Y, et al., 2012 ELK Reasoner: Architecture and Evaluation. In: Horrocks I, et al., (Eds.) Proceedings of the OWL Reasoner Evaluation Workshop 2012 (ORE’12), vol. 858. CEUR Workshop Proceedings, CEUR-WS.org.

Köhler S, 2013 Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. F1000Res. doi: 10.12688/f1000research.2-30.v2.

Lin D, 1998 An Information-Theoretic Definition of Similarity. In 15th International Conference of Machine Learning Madison, WI: 1998:296–304.

McMurry JA, et al., 2016 Navigating the Phenotype Frontier: The Monarch Initiative. Genetics 203, 1491–1495. [PubMed: 27516611]

Mendez J, 2012 jcel: A Modular Rule-based Reasoner. In Proceedings of the 1st International Workshop on OWL Reasoner Evaluation (ORE 2012), 858.

Merrick B, 2015 Intersection of toxicogenomics and high throughput screening in the Tox21 program: an NIEHS perspective. Int. J. Biotechnol 14(1), 7–27. [PubMed: 27122658]

Mungall CJ, et al., 2010 Integrating phenotype ontologies across multiple species. Genome Biol 11(1), R2. doi: 10.1186/gb-2010-11-1-r2. [PubMed: 20064205]

National Research Council, 2007 Toxicity Testing in the 21st Century: A Vision and a Strategy. Washington: The National Academies Press.

Nilsson S, et al., 2001 Mechanisms of Estrogen Action. Physiol. Rev 81, 1536–1565.

Osumi-Sutherland D, et al., 2017 Dead simple OWL design patterns. J. Biomed. Semantics, 8:18. [PubMed: 28583177]

Pape-lindstrom PA, Lydy MJ, 1997 Synergistic toxicity of atrazine and organophosphate insecticides contravenes the response addition mixture model. Environ. Toxicol. Chemi 16(11), 2415–2420.

Pesquita C, et al., 2007 Evaluating GO-based semantic similarity measures. Proc. 10th Annual Bio- 2007:1–4.

Pesquita C, et al., 2008 Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9(Suppl 5), S4. doi:10.1186/1471-2105-9-S5-S4.

Pesquita C, et al., 2009 Semantic Similarity in Biomedical Ontologies. PLoS Comput. Biol 5(7), e1000443. doi:10.1371/journal.pcbi.1000443. [PubMed: 19649320]

Richard A, et al., 2016 ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chem Res Toxicol 29(8), 1225–51. doi: 10.1021/acs.chemrestox.6b00135. [PubMed: 27367298]

Sanchez D, et al., 2011 Ontology-based information content computation. Knowl. Based Syst 24:297–303.

Shimabukuro RH, Swanson HR, 1969 Atrazine metabolism, selectivity, and mode of action. J. Agric. Food Chem 17 (2), 199–205. doi: 10.1021/jf60162a044.

Smith B, et al., 2007 The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11), 1251–1255. doi: 10.1038/nbt1346. [PubMed: 17989687]

Washington NL, et al., 2009 Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation. PLoS Biol 7(11), e1000247. doi:10.1371/journal.pbio.1000247. [PubMed: 19956802]

Xiang Z, et al., 2010 OntoFox: web-based support for ontology reuse. BMC Res. Notes 3, 175. [PubMed: 20569493]

Yu G, et al., 2010 “GOSemSim: a R package for measuring semantic similarity among GO terms and gene products.” Bioinformatics 26(7), 976–978. doi: 10.1093/bioinformatics/btq064. [PubMed: 20179076]

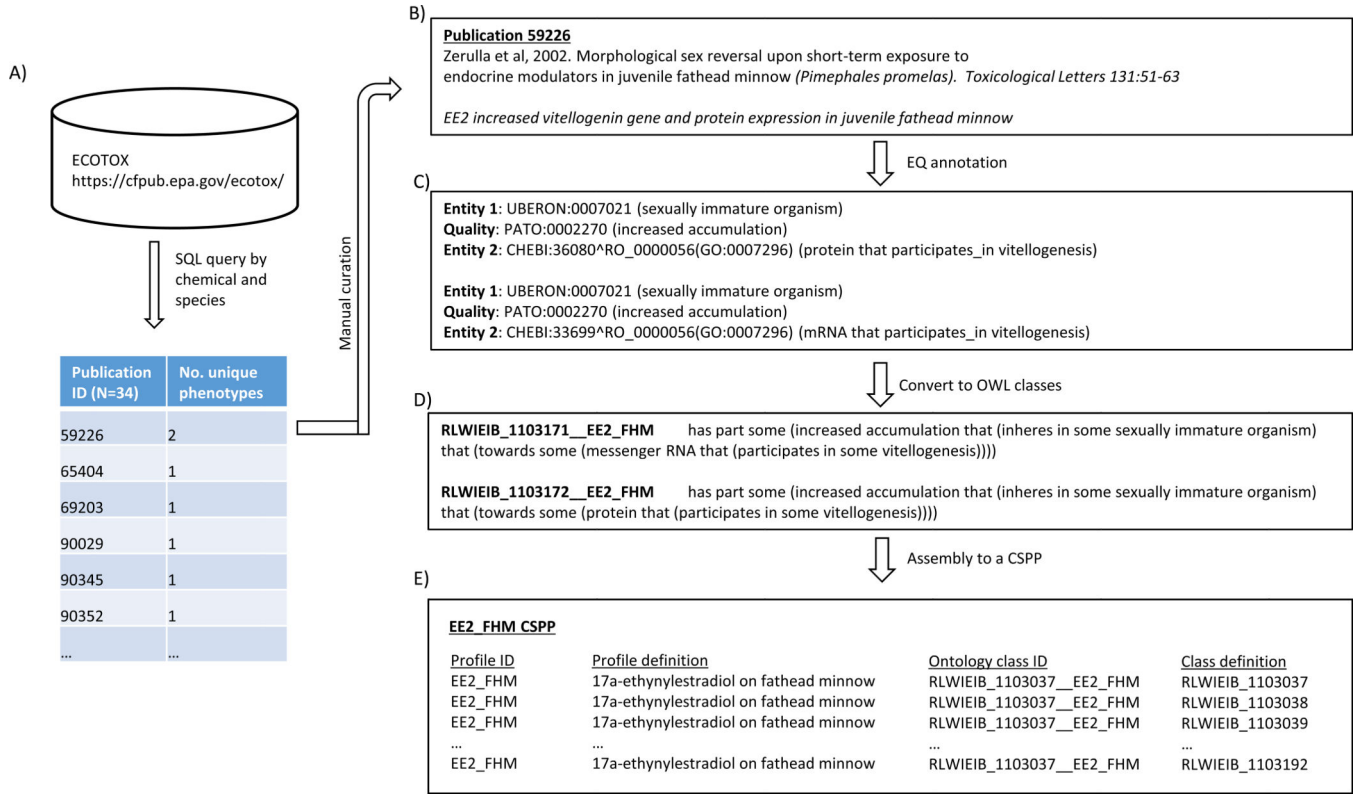
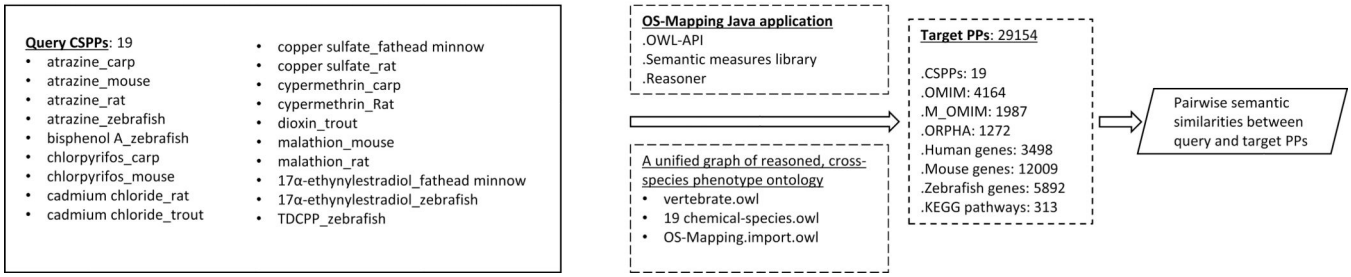


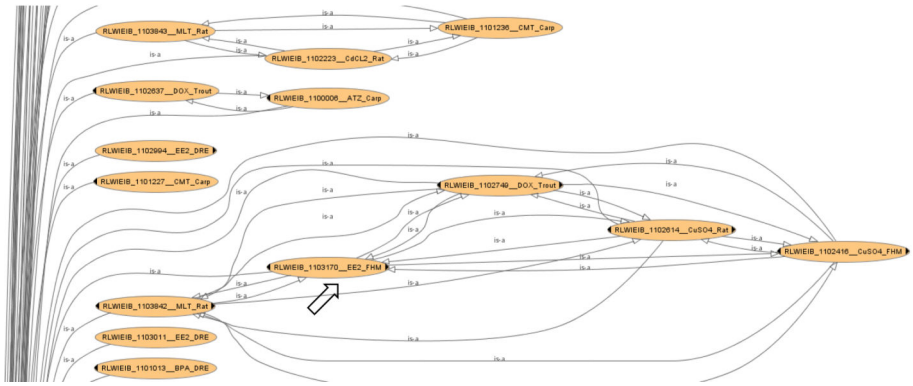
Figure 1.

The workflow to build a chemical-species phenotypic profile (CSPP) as illustrated by the 17α-ethynylestradiol_fathead minnow (EE2_FHM). A) Relevant publications were first identified from ECOTOX; B) two 17α-ethynylestradiol-induced phenotypes reported in the publication 59226 were curated manually; C) the phenotypes were annotated in Entity-Quality (EQ) syntax; D) EQ annotations were converted to their respective ontology classes in Manchester syntax; E) all annotated ontology classes for the 17α-ethynylestradiol _fathead minnow were organized into its CSPP.

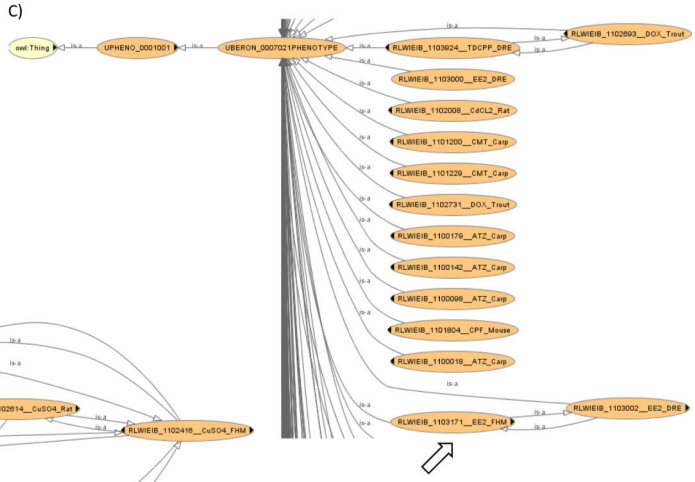
A)



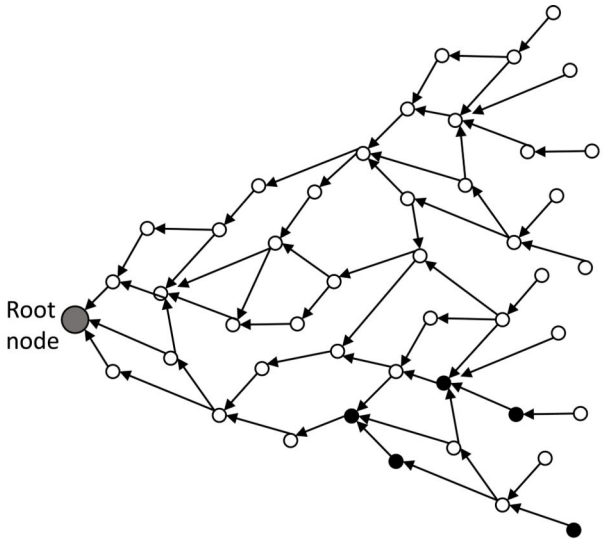
B)



C)



D)



E)

query PP	target PP	similarity	query PP size	target PP size	target PP definition	cutoff P0.05	cutoff P0.01
EE2_FHM	EE2_FHM	1.000	156	156	17a-ethynylestradiol on fathead minnow	0.500	0.512
EE2_FHM	M_Ccnb1ip1	0.704	156	14	cyclin B1 interacting protein 1	0.569	0.590
EE2_FHM	M_repro15	0.698	156	4	reproductive mutant 15	0.571	0.590
EE2_FHM	M_Dms	0.696	156	7	dominant male sterility	0.588	0.604
EE2_FHM	M_Cdk16	0.695	156	8	cyclin-dependent kinase 16	0.597	0.609
EE2_FHM	M_Cnot7	0.694	156	20	CCR4-NOT transcription complex, subunit 7	0.593	0.610
EE2_FHM	EE2_DRE	0.694	156	269	17a-ethynylestradiol on zebrafish	0.522	0.538
...	...	...	...	...	...	...	...

Figure 2.

OS-Mapping analysis workflow. A) OS-Mapping java application compared the query against target phenotypic profiles semantically based on vertebrate.owl, a cross-species phenotype ontology. Prior to the analysis, vertebrate.owl first imported many other domain ontologies, and then merged them with both the custom OWL (Web Ontology Language) classes constructed from the Entity-Quality annotations of various chemical-species ('19 chemical-species.owl') and the external ontology terms contained therein (OS-Mapping.import.owl). The merged ontology was subsequently reasoned into a unified ontology graph, based on which CSPPs were analyzed; B, C) two sample subgraphs containing two of the 156 ontology classes of 17α-ethynylestradiol_fathead minnow from the merged and reasoned ontology as extracted by the robot tool (http://robot.obolibrary.org); D) an illustrative ontology graph with classes represented by nodes and their subsumption relationships by edges. A query profile (solid black nodes) and target profiles (various combinations of nodes in the entire graph) are compared semantically; E) part of a simplified 17α-ethynylestradiol_fathead minnow query output.

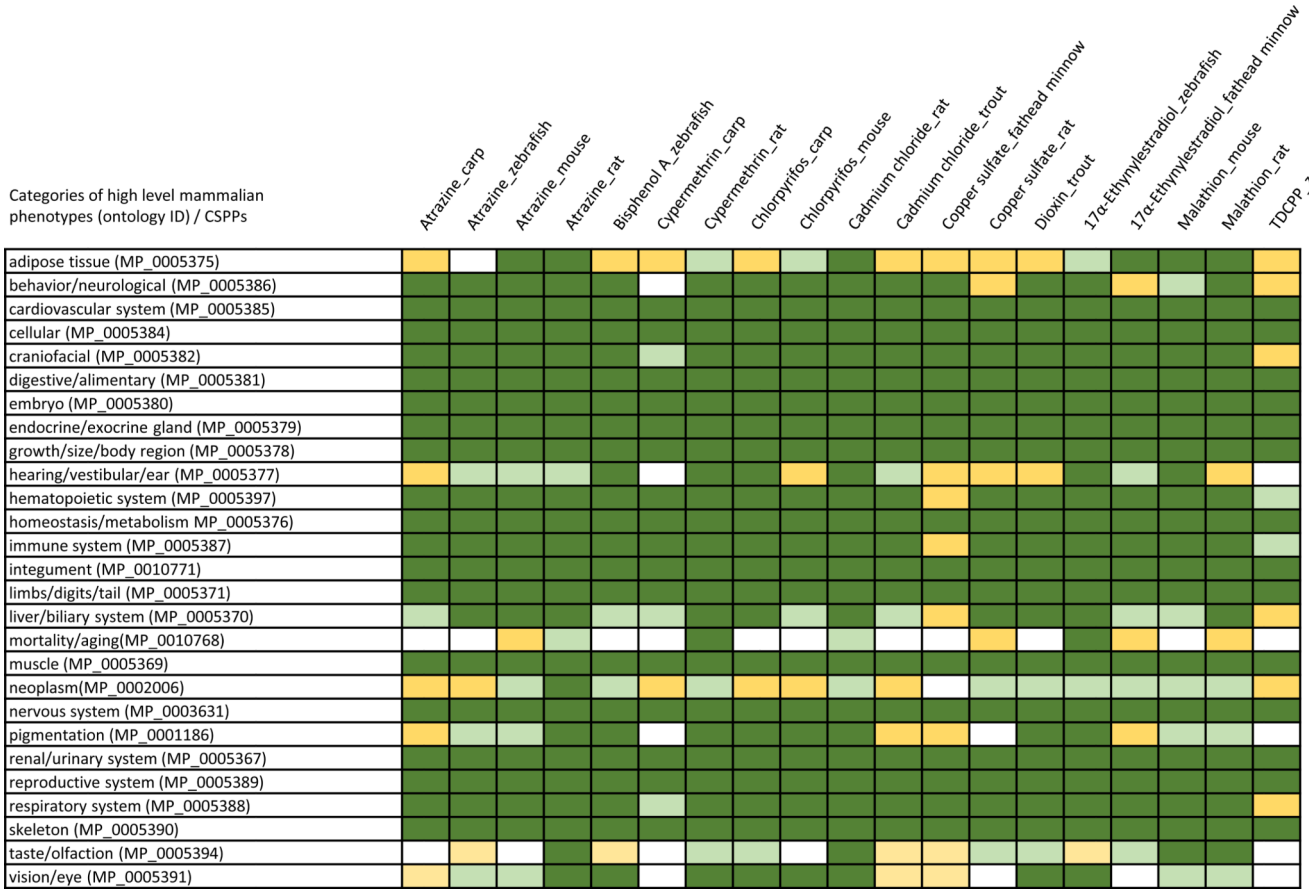


Figure 3.

The scope of coverage of mammalian phenotypes by chemical-species phenotypic profiles (CSPPs). The Mammalian Phenotype Ontology (MP) terms with pairwise similarities to the custom ontology classes from individual CSPPs ≥ 0.9, 0.8, or 0.7 were retained for extracting their respective superclasses in mp.owl using the Robot tool (http://robot.obolibrary.org). The 27 categories of high level MP terms (http://www.informatics.jax.org) present in each CSPP-associated MP superclasses based on the minimum similarity scores of 0.9, 0.8, or 0.7 are marked as dark green, light green, or yellow respectively. The categories absent at ≥ 0.7 are marked as white. The phenotypic coverages at these minimum scores are 77% (393/27×19), 85% (438/27×19), and 95% (485/27×19). TDCPP, Tris(1,3-dichloroisopropyl) phosphate.

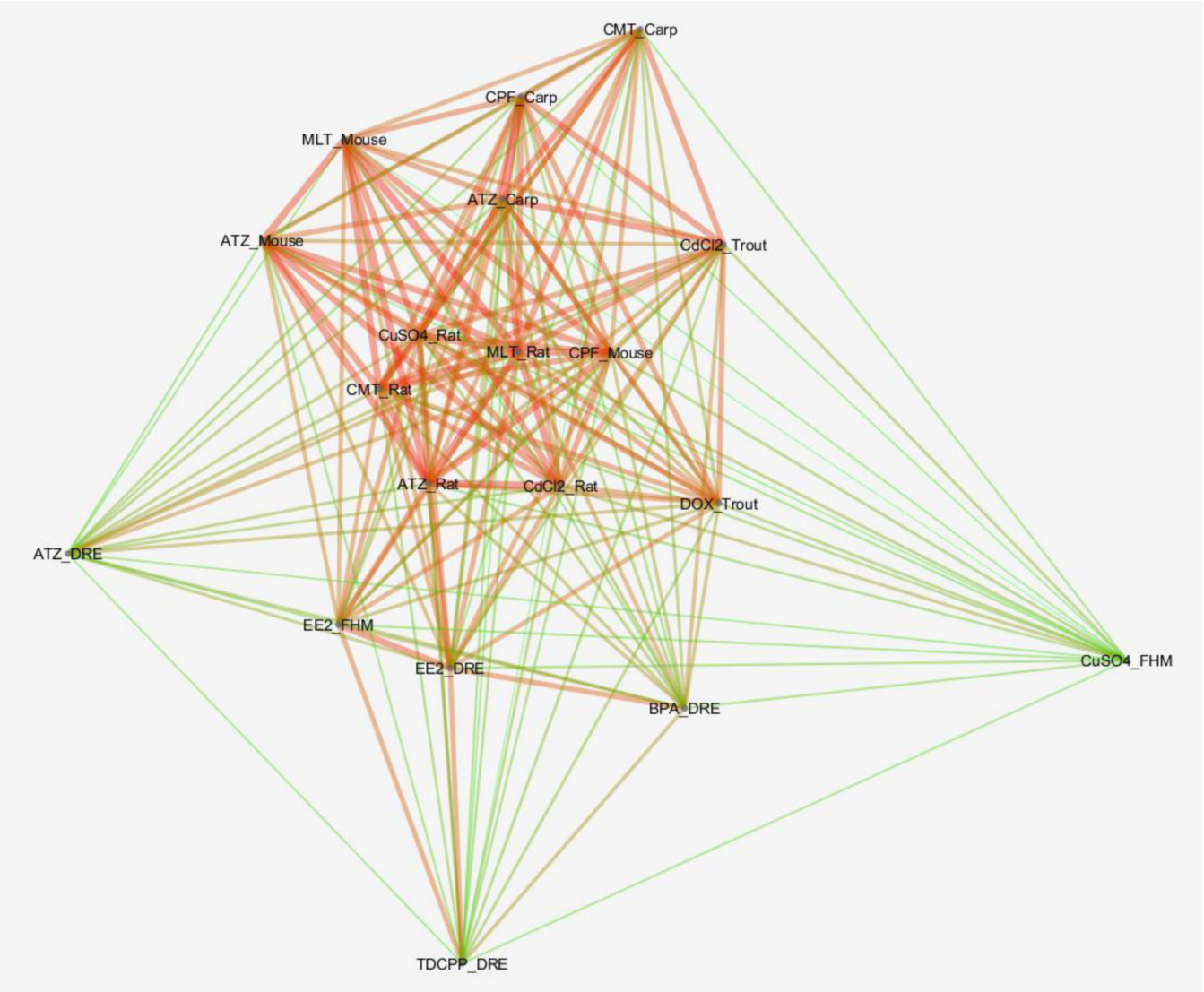


Figure 4.

Semantic similarity network of chemical-species phenotypic profiles (CSPPs). Edges are weighted by indirect group-wise similarity scores ranging between 0 and 1, in “prefuse force directed layout” at 3X scale in Cytoscape (http://www.cytoscape.org), with shorter and wider edges in darker red color indicating a greater similarity between the two connected nodes. Abbreviations: ATZ, atrazine; BPA, bisphenol A; CdCl2, cadmium chloride; CMT, cypermethrin; CPF, chlorpyrifos; CuSO4, copper sulfate; DOX, dioxin; DRE, zebrafish; EE2, 17α-Ethynylestradiol; FHM, fathead minnow; MLT, malathion; TDCPP, Tris(1,3-dichloroisopropyl) phosphate.

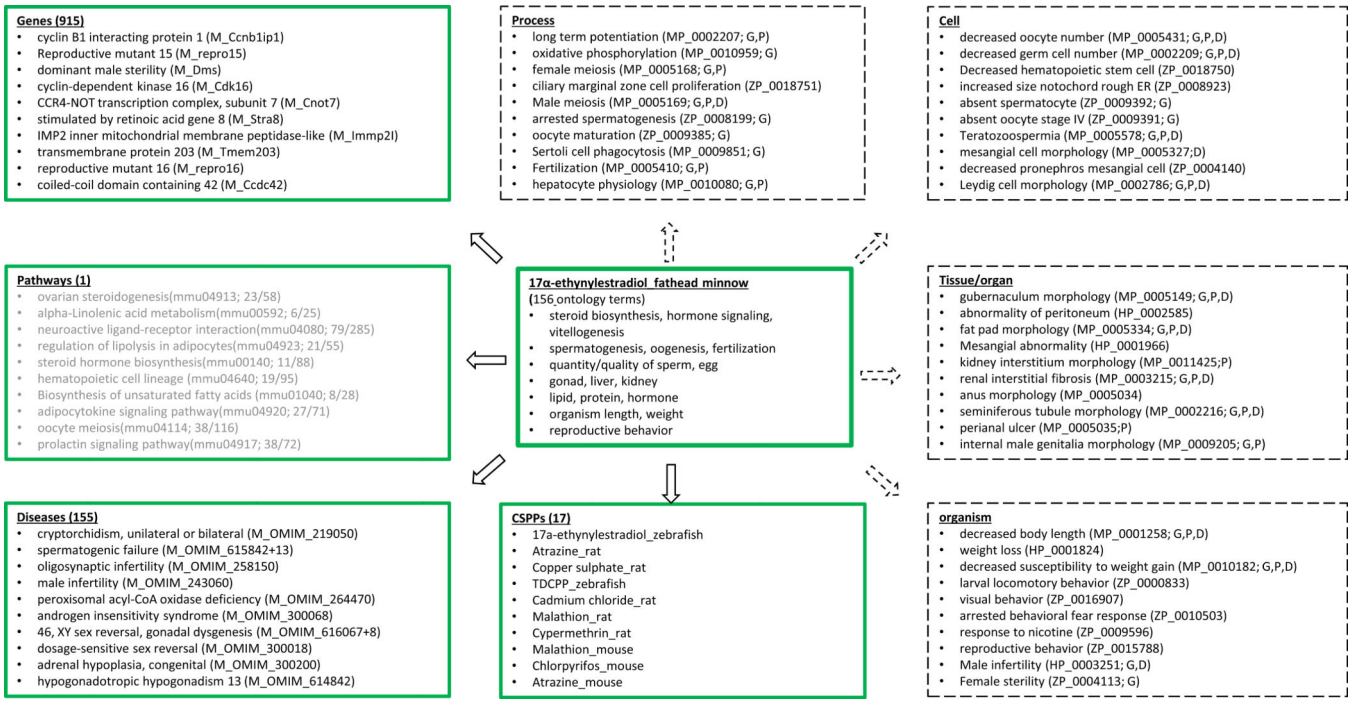


Figure 5.

A phenotypic continuum anchored by the 17α-ethynylestradiol_fathead minnow. Top ten mapped phenotypic profiles of genes ($P_{0.01}$), pathways, CSPPs ($P_{0.01}$), and disease ($P_{0.01}$) are listed. The total counts of genes, pathways, and diseases are based on $P_{0.05}$. The top ten KEGG pathways listed by OS-Mapping are not significant at $P_{0.05}$ (text in gray). Each pathway is denoted in a parenthesis by its KEGG ID, number of genes whose mutation/alleles directly caused phenotypes in reproductive systems (MP_0005389), and total number of genes in the pathway. The phenotypes at the levels of biological process, cell, tissue/organ, and organism (boxes in dash lines) were selected from the top five HP/MP/ZP terms present in the target phenotypic profiles, based on their semantic similarities to individual terms in the 17α-ethynylestradiol_fathead minnow. The terms overlapping with the phenotypic profiles of 915 genes, the top ten pathways, and 155 diseases are marked by G, P, and D in parentheses. A disease with multiple variants are denoted by its OMIM ID + the number of variants, for example, M_OMIM_615842+13. M_OMIM, an OMIM disease annotated by MP terms. Abbreviations: KEGG, Kyoto Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg); OMIM, Online Mendelian Inheritance in Man (https://www.omim.org).

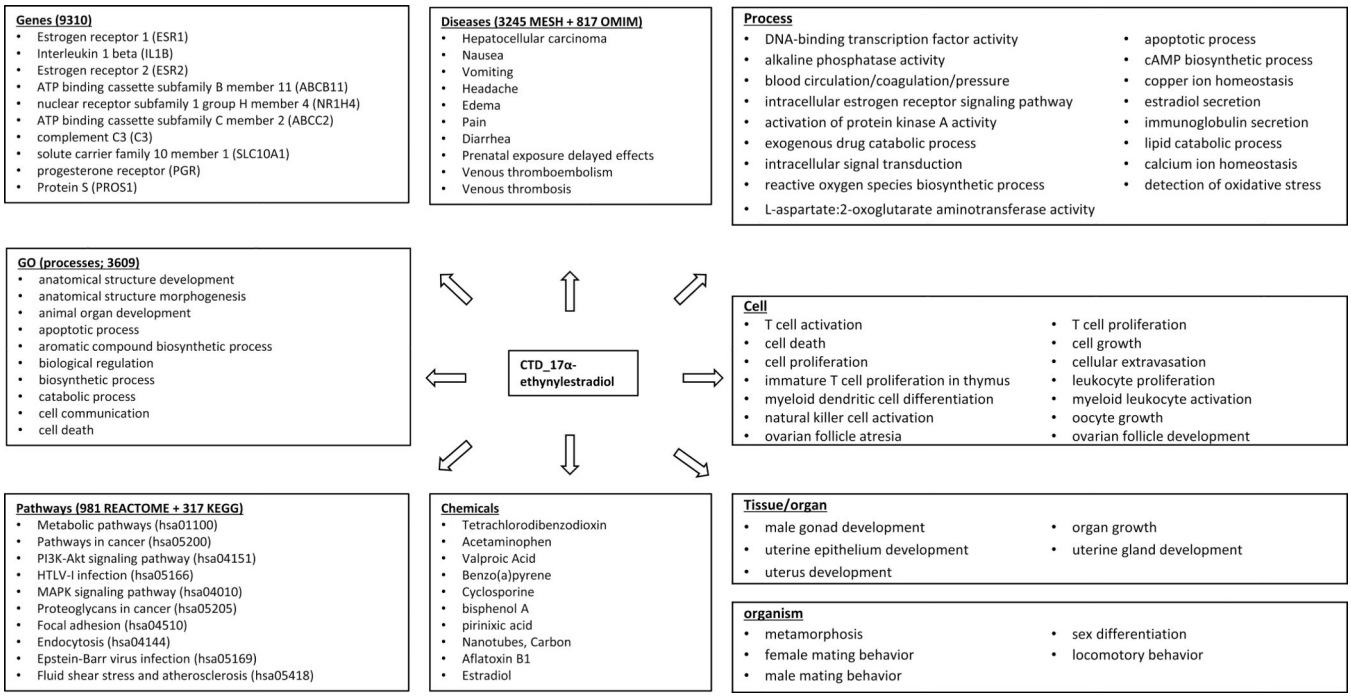


Figure 6.

Mappings of 17α-ethynylestradiol according to the Comparative Toxicogenomics Database (http://ctdbase.org; as of July 2018). The mappings were constructed as follows: chemical-genes, literature curation; chemical-pathways, human KEGG (hsa) and REACTOME pathways enriched in chemical-genes; chemical-Gene Ontology (GO) terms, GO terms enriched in chemical-genes; chemical-disease, direct curation from the literature (4.5%) or indirect inference from transitive chemical-gene-disease associations (99%), with a slight overlap between the two methods; chemical-phenotypes (delineated by process, cellular, tissue/organ, and organism), literature curation; chemical-chemical linkages, shared gene interactions. Except for phenotypes, only top ten mappings are listed for each category. For pathways, only human KEGGs are listed.

Table 1.

Summary of semantic mapping hits by chemical-species phenotypic profiles (CSPPs; P-value ≤ 0.05). Significant hits are the row sums of mapped CSPPs, genes, pathways, and diseases for individual CSPPs. Pearson correlation: r (profile size, significant hits) = 0.41, r (phenotypic coverage, significant hits) = 0.56.

CSPP (No. publications)	Profile size[1]	Phenotypic coverage[2]	CSPP[3]	Mouse Genes	Human Genes	Zebrafish Genes	KEGG pathways	Disease accessions[4]	Significant hits
cypermethrin _rat (40)	200	1.00	19	2059	201	102	50	549	2980
malathion _mouse (49)	182	0.96	18	1901	127	138	66	440	2690
atrazine _mouse (28)	187	0.96	18	1564	126	82	61	478	2329
atrazine _rat (94)	477	1.00	19	1275	110	55	31	337	1827
malathion _rat (93)	480	1.00	19	1248	85	73	39	289	1753
cadmium chloride_trout (30)	130	0.96	19	885	89	51	8	258	1310
cadmium chloride_rat (76)	404	1.00	19	898	47	7	18	169	1158
17α-ethynylestradiol_fathead minnow (34)	156	0.96	17	800	66	49	1	155	1088
17α-ethynylestradiol_zebrafish (66)	269	1.00	18	757	73	52	1	165	1066
copper sulfate_rat (28)	193	0.93	19	596	54	8	2	169	848
dioxin_trout (26)	139	0.96	19	375	84	45	4	226	753
chlorpyrifos _carp (27)	216	0.96	19	414	35	17	0	175	660
atrazine _carp (31)	244	0.93	19	354	21	17	0	120	531
cypermethrin _carp (12)	96	0.78	17	199	21	21	0	57	315
chlorpyrifos _mouse (42)	187	0.93	19	220	4	3	1	25	272
bisphenol A_zebrafish (26)	132	0.96	16	40	0	30	0	0	86
Copper sulfate_fathead minnow (23)	58	0.93	13	52	0	8	0	0	73
atrazine _zebrafish (22)	105	0.93	15	18	0	5	0	0	38
Tris(1,3-dichloroisopropyl) phosphate _zebrafish (8)	144	0.81	9	15	0	2	0	0	26
Total unique (726)	3999	---	19	3702	422	338	94	1181	5756

[1] the number of post-composed OWL classes included in a CSPP.

[2] percentage (×100) of the 27 categories of high level phenotypes represented (http://www.informatics.jax.org).

[3] including self-hits.

[4] including human diseases (OMIM; https://www.omim.org) annotated by either HP or MP terms, and rare diseases (Orphanet; https://www.orpha.net/consor/cgi-bin/index.php).