D948–D954 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 24 September 2018*
doi: 10.1093/nar/gky868

The Comparative Toxicogenomics Database: update 2019

Allan Peter Davis[1,*], Cynthia J. Grondin[1], Robin J. Johnson[1], Daniela Sciaky[1], Roy McMorran[2], Jolene Wiegers[1], Thomas C. Wiegers[1] and Carolyn J. Mattingly[1,3]

[1]Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695, USA, [2]Department of Bioinformatics, The Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672, USA and [3]Center for Human Health and the Environment, North Carolina State University, Raleigh, NC 27695, USA

Received August 14, 2018; Revised September 10, 2018; Editorial Decision September 12, 2018; Accepted September 14, 2018

ABSTRACT

The Comparative Toxicogenomics Database (CTD; http://ctdbase.org/) is a premier public resource for literature-based, manually curated associations between chemicals, gene products, phenotypes, diseases, and environmental exposures. In this biennial update, we present our new chemical–phenotype module that codes chemical-induced effects on phenotypes, curated using controlled vocabularies for chemicals, phenotypes, taxa, and anatomical descriptors; this module provides unique opportunities to explore cellular and system-level phenotypes of the pre-disease state and allows users to construct predictive adverse outcome pathways (linking chemical–gene molecular initiating events with phenotypic key events, diseases, and population-level health outcomes). We also report a 46% increase in CTD manually curated content, which when integrated with other datasets yields more than 38 million toxicogenomic relationships. We describe new querying and display features for our enhanced chemical–exposure science module, providing greater scope of content and utility. As well, we discuss an updated MEDIC disease vocabulary with over 1700 new terms and accession identifiers. To accommodate these increases in data content and functionality, CTD has upgraded its computational infrastructure. These updates continue to improve CTD and help inform new testable hypotheses about the etiology and mechanisms underlying environmentally influenced diseases.

INTRODUCTION

The Comparative Toxicogenomics Database (CTD; http://ctdbase.org/) is a premier public resource that advances understanding about chemical exposures and human health (1–5). We use novel biocuration paradigms (6) to curate disparate data for toxicogenomics (7–10), phenotypes (11–12), diseases (11,13–16), environmental exposures (17–20), and pharmaceuticals (11,21). Professional biocurators manually curate the scientific literature, transforming text, tables, figures and supplemental files into annotated data that is seamlessly integrated and available through CTD's public web application (PWA). This process, using a suite of community-accepted controlled vocabularies and ontologies with accession identifiers, ensures CTD content is cohesive, manageable, and computable, as well as adhering to the FAIR principle, allowing the information to be Findable, Accessible, Interoperable and Reusable (22). Additionally, all CTD interactions are embellished with taxa identification (enabling data to be compared across species, from model laboratory organisms to humans) and are directly linked to the original source article (providing full transparency and traceability). Furthermore, CTD maintains good data stewardship with the digital informatics community by facilitating semantic standards for the environmental health science community (23), complying with reporting standards set by the BioSharing Information Resources (24), and registering with BioDBcore (https://biosharing.org/biodbcore-000173) (25) and the NAR Molecular Biology Database Collection (http://www.oxfordjournals.org/our_journals/nar/database/summary/1188).

Here, we provide our biennial database update and describe our newly released chemical–phenotype module, advanced query and display tools for exposure science data, an extensive update to the MEDIC disease vocabulary and an overall increase in manually curated content. Researchers

*To whom correspondence should be addressed. Tel: +1 919 515 5705; Fax: +1 919 515 3355; Email: apdavis3@ncsu.edu

Table 1. Updated CTD content (August 2018)

Data type	2018 Counts
Scientific articles	129 564
Chemicals	15 681
Genes	46 689
Diseases	7212
Phenotypes	4340
Anatomy	799
Taxa	583
chemical–gene interactions	1 812 207
gene–disease interactions	37 782
chemical–disease interactions	211 974
Phenotype-based interactions	169 697
Exposure statements	122 725
Sub-total (manually curated interactions)	2 354 385
gene–disease inferences	24 261 740
chemical–disease inferences	2 295 455
Chemical-GO inferences	5 180 064
Chemical-pathway inferences	1 179 283
Disease-pathway inferences	283 403
Disease-GO inferences	938 758
Imported gene-GO annotations	1 212 332
Imported gene-pathway annotations	135 805
Imported gene-gene interactions	503 343
Total	38 344 568

can leverage CTD to explore novel connections and quickly generate testable hypotheses about the molecular mechanisms of chemical influenced health outcomes.

NEW FEATURES

Increased data content

It is prudent for databases to update their content routinely to stay relevant and to keep pace with scientific advances. As of August 2018, CTD includes over 2.3 million manually curated chemical–gene, chemical–phenotype, chemical–disease, gene–disease and chemical–exposure interactions for 15 681 chemicals, 46 689 genes, 4340 phenotypes and 7212 diseases (Table 1), representing a 46% increase in content since our last update (5). These interactions are manually curated from 129 564 peer-reviewed scientific articles, triaged from PubMed using targeted journal queries to enhance data currency (26) and chemical-centric queries and text-mining to improve data completeness and increase curatorial efficiency and productivity (27). This workflow keeps CTD both relevant and up-to-date with the toxicology and exposure literature, as new content is added every month (http://ctdbase.org/about/dataStatus.go); as well, CTD's controlled vocabularies and curated content are described and made freely available for users to download in a variety of formats (.csv, .obo, .tsv and .xml) at 'Data Downloads' (http://ctdbase.org/downloads/).

Internal integration of these direct interactions (9) generates >24 million gene–disease and 2.2 million chemical–disease predictive inferences that are statistically ranked (28). External integration of CTD content with annotations from Gene Ontology (29), KEGG (30), Reactome (31) and BioGRID (32) produces an additional 9.4 million inferences (Table 1). In total, CTD includes over 38 million toxicogenomic relationships for analysis and hypothesis development, and was recently named a 'golden set database' that serves as an 'authoritative, comprehensive, and convenient data resource' (33). This repute is corroborated by >1500 citations to CTD (including over 730 since 2016) and by 100 external databases now incorporating or linking to CTD content (http://ctdbase.org/about/publications/#use).

New CTD chemical–phenotype module

While CTD has been curating chemical-induced diseases since 2007 (1), many toxicology articles do not report a disease as an endpoint, but rather describe molecular, cellular, or physiological systems that are altered, such as signal transduction, cell proliferation, apoptosis, and immune system processes. To curate these important non-disease events, we developed a new module that codes chemical-induced phenotypes (12). At CTD, we operationally distinguish phenotypes from diseases: if the reported outcome exists in CTD's MEDIC disease vocabulary (15), we consider it a disease; subsequently, any outcome not in MEDIC is considered de facto a phenotype and is coded in our new module. CTD chemical-induced phenotypes are curated in a structured format using controlled vocabularies and include six components (chemical, action qualifier, phenotype entity, taxon, anatomy and PubMed reference) as well as an inference network (Figure 1). For phenotype curation, we co-opted the Gene Ontology (GO) as a vocabulary source for non-disease biological outcomes; this has proven remarkably successful since most of the phenotypes reported in the toxicology literature already exist as a controlled term in the GO, and GO is a well-known, community-accepted vocabulary with accession identifiers, allowing novel CTD chemical–phenotype interactions to be computable and interoperable with other databases. Additionally, CTD chemical–phenotype interactions are enhanced with an organism term from NCBI Taxonomy (34) and anatomical descriptions from Medical Subject Headings (MeSH) 'Anatomy' [A] branch (35).

Chemical–phenotype interactions are reciprocally displayed on CTD Chemical pages (under the 'Phenotypes' data-tab; Figure 1) and on CTD GO pages (under the 'Chemical Interactions' data-tab). The Inference Network automatically computes a list of genes that have both a directly curated chemical–gene interaction in CTD and are independently annotated to the same GO term by external databases (36,37). The Inference Network identifies potential molecular networks that connect the chemical to the phenotype by finding common genes from two completely independent curated datasets. Users can find a quick guide about the phenotype module by clicking the *Help* icon ('?') on the webpage (Figure 1).

CTD chemical–phenotype interactions provide numerous discovery opportunities for users. They allow phenotypes to be explored easily from a chemical perspective (e.g. bisphenol A) or a phenotype perspective (e.g. apoptosis), providing insight into mechanistic connections across species and anatomical tissues (Figure 2A). As well, by connecting accession identifiers for hierarchical terms to both the CTD chemical environment and GO phenotypes (Figure 2B), these interactions help address the community's need to link phenotypes to the environment and make the information computable for meta-analyses and discovery (37). Third, CTD chemical–phenotype data complements

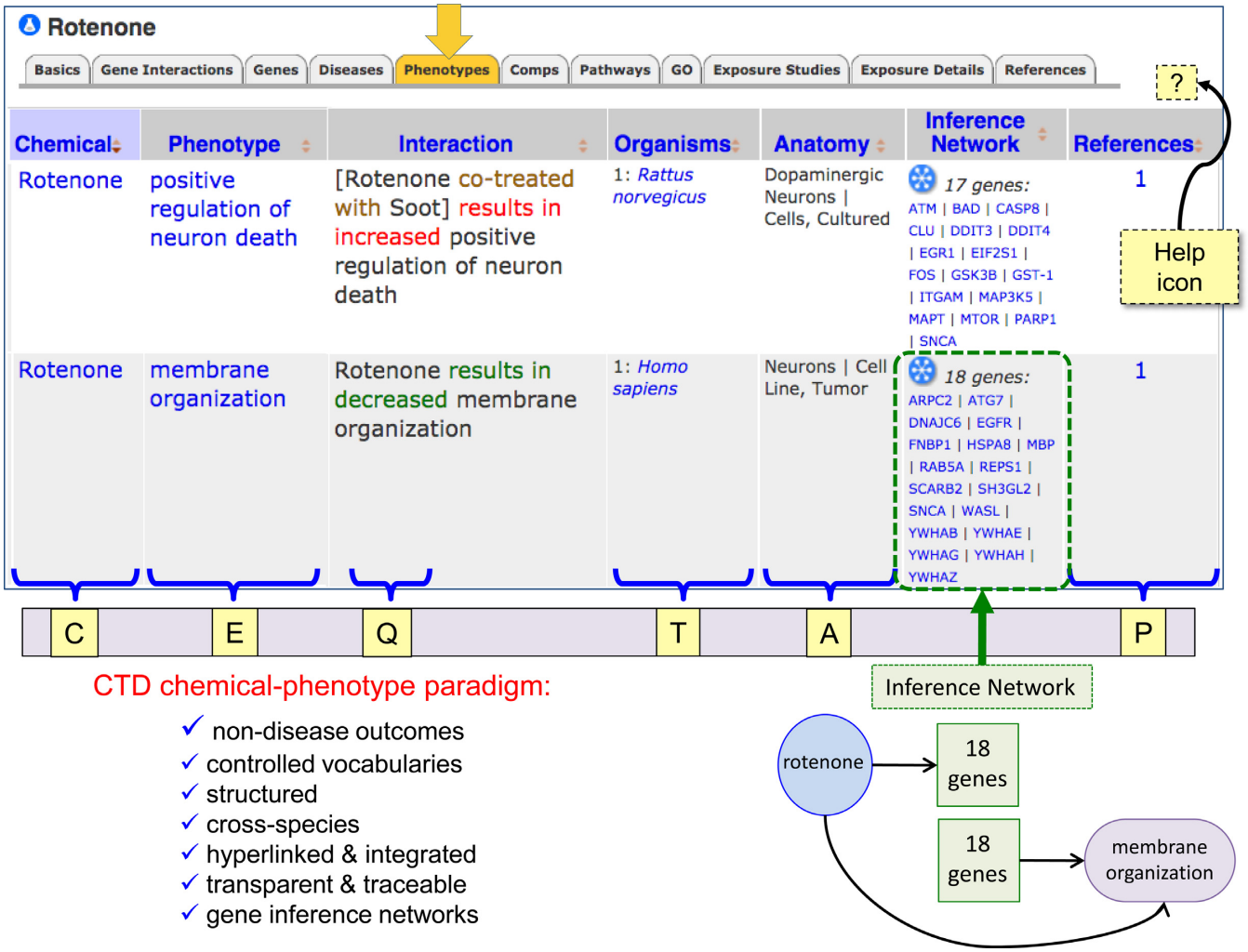

Figure 1. CTD's new phenotype module. CTD's chemical–phenotype curation paradigm collects novel information for chemically induced non-disease outcomes across species, with hyperlinked terms to allow seamless navigation across CTD. chemical–phenotype interactions are displayed under the new 'Phenotypes' data-tab on chemical pages (and vice versa under the new 'Chemical Interactions' data-tab on GO/Phenotype pages). Interactions are curated in a structured format using controlled vocabularies for chemicals (C), phenotype entities (E), action qualifiers (Q, 'increased', 'decreased', or 'affects'), organisms (T) and anatomy (A), and are directly traceable to the source article (P). Additionally, *Inference Networks* list a set of genes that provide a putative molecular framework to connect the chemical to the phenotype. Here, the insecticide rotenone affects several phenotypes, including 'membrane organization' in a human neural tumor cell line; as well, rotenone directly interacts with 18 genes in CTD that are also independently annotated to the same GO term, forming an inference network. The *Help* icon ('?') provides users with a link to a concise guide about the phenotype module. For simplicity, an edited screenshot is shown.

the myriad of well-established gene–phenotype modules from other databases, and integration (via numerous shared accession IDs) enables model organism gene-phenotype information to be brought into the chemical exposure environment provided by CTD (Figure 2C). Furthermore, combining CTD's chemical–phenotype data with CTD's chemical–disease data allows phenotypes to be inferred to diseases (via shared chemicals) to provide insight into potentially presymptomatic conditions of a disease (Figure 2D). This integration connects sub-cellular phenotypes shared by different diseases, potentially informing earlier diagnosis and new therapeutic strategies (16). Finally, integrating all four CTD curation modules can be used to compute predictive adverse outcome pathways (AOP), which depict toxicant-induced events affecting human health in a series of interoperable modular graphs (38). CTD's four curation projects mirror the four components of an AOP: CTD's toxicogenomic core includes the chemical–gene interactions that correspond to molecular initiating events (MIE), CTD's chemical–phenotype interactions describe key events (KE) in the pathway, CTD's disease core reports chemical- and gene-induced adverse outcomes (AO), and our exposure module relates exposure stressor-induced effects at the population level (PO). Integrating CTD content can help quickly generate predictive AOPs for analysis, information discovery, and testing (Figure 2E), as we demonstrated in building predictive AOPs describing arsenic-mediated MIEs for 39 genes affecting ten glucose-related metabolic KEs prior to the onset of a diabetic AO, as well as ten predictive AOPs for cadmium-induced phenotypes (cell signaling cascades, neuronal apoptosis and al-

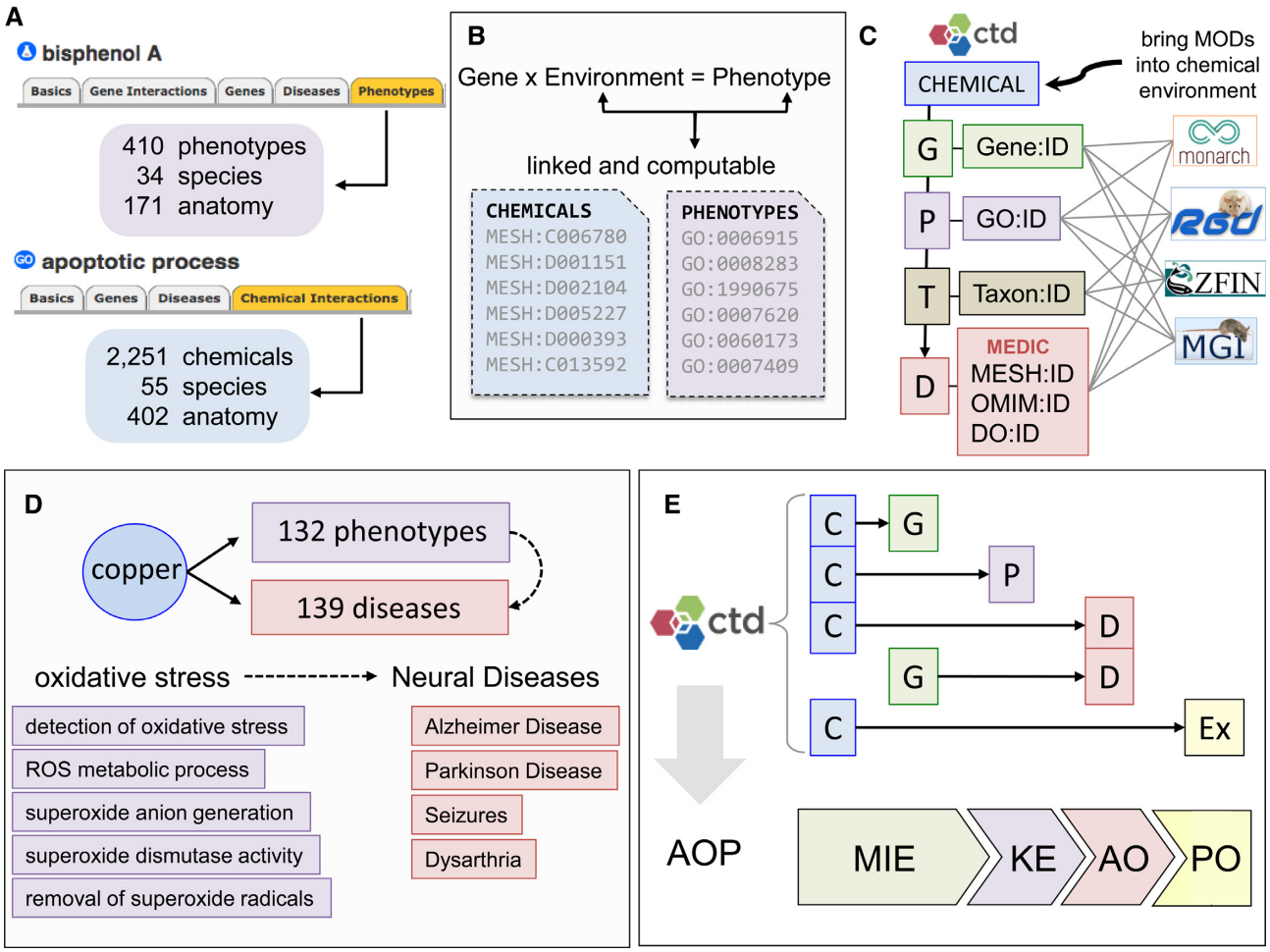

Figure 2. Using chemical–phenotype data for discovery. (**A**) Users can explore phenotypes from a chemical perspective (e.g. bisphenol A influences 410 phenotypes in 171 anatomical sites from 34 species) as well as discover chemicals that affect a specific phenotype (e.g. apoptotic process is modulated by 2,251 chemicals in 55 species). (**B**) Use of hierarchical controlled vocabularies (with accession identifiers) for both chemicals (MESH:ID) and phenotypes (GO:ID) provides data files that allow environmental factors and phenotypes to be linked and computable for meta-analyses. (**C**) Use of interoperable accession identifiers for genes (G), phenotypes (P), taxa (T), and diseases (D) shared by model organism databases (MODs) and other resources enable their content to be integrated with and brought into the chemical environment of CTD. (**D**) Non-disease phenotypes can be inferred to diseases (based on shared interacting chemicals) to help inform the pre-disease state. The heavy metal copper modulates 132 phenotypes and, independently, is associated with 139 diseases in CTD, providing a view of the potential biological processes in the presymptomatic condition, such as the numerous oxidative stress phenotypes that might precede the onset of neurological diseases. This knowledge can be leveraged to find novel commonalities between sets of phenotypes and diverse diseases, with the potential of re-purposing or discovering new therapeutic drugs. (**E**) Integrating data from all four CTD modules helps generate predictive adverse outcome pathways (AOP). CTD's toxicogenomic core reports chemical–gene (C–G) interactions that parallel the molecular initiating event (MIE) of an AOP, CTD's new phenotype module links chemical–non-disease phenotype (C–P) key events (KE), disease core curates chemical–disease (C–D) and gene–disease (G–D) adverse outcomes (AO), and CTD's exposure module relates chemical exposures (C-Ex) for population-level health outcomes (PO).

tered learning and memory) that may precede Alzheimer disease (12).

chemical–phenotype interactions can also be explored using a new *chemical–phenotype Interaction Query* (http://ctdbase.org/query.go?type=phenotype). This allows users to perform advanced searches based not only on chemicals and phenotypes, but also the hierarchical controlled terms for anatomy and organisms. Thus, a researcher interested in nephrotoxicity can search the Anatomy field with the word 'kidney' to retrieve over 9,300 interactions for more than 1100 chemicals affecting 510 phenotypes in 17 renal structures from 30 species (http://bit.ly/ctdkidneyphenotype).

Increased content, utility, and functionality of exposure science data

In 2015, CTD launched the first comprehensive, literature-based, manually curated exposure science module that centralized and harmonized environmental science data from diverse research studies by using controlled vocabularies for chemical stressors, receptor demographics, biomarkers, exposure outcomes, and geographical locations, among other data types (18,19). This module is fully integrated within the CTD framework, connecting real-world measurements for environmental toxicants and biomarkers with laboratory-derived toxicogenomic data. Since our last update, we have improved the utility of this module in three important ways.

First, as mentioned (Table 1), the content has dramatically increased to 122 725 chemical–exposure statements (74% increase). These statements are curated from 2150 articles and relate exposure data for 1155 chemical stressors, 432 human genes, 421 environmental diseases and 332 phenotypes.

Second, we enhanced the functionality of the *Exposure Studies* query page (http://ctdbase.org/query.go?type=expStudies) by including two additional search parameters. A user can now add *Study Factors* to screen for 11 entities affecting the conclusions of an exposure study, such as diet, age, race, socioeconomic status, *etc*. The additional query field *Associated Study Titles* allows selection of curated articles from a common cohort, with over 415 projects alphabetically listed (from the Aarhus Birth Cohort of Denmark to the Wuhan-Zhuhai Cohort of China). In exposure science, large, cooperative, multi-year projects and datasets are often given a study title and acronym. The integration of study titles helps users to identify and collate articles derived from a project (often published over many years by different groups in various journals), and enables users to conduct meta-analyses across many publications, as CTD previously demonstrated with the Agriculture Health Study (18). Data from the National Health and Nutrition Examination Survey (NHANES) are extensively used in diverse exposure analyses. Currently, CTD has over 175 articles curated with NHANES-related data, and users can quickly peruse them by using the *Associated Study Titles* search parameter (http://bit.ly/ctdnhanes). Alternatively, a researcher interested in the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) can retrieve the many articles reporting data from this birth cohort study of pesticides (http://bit.ly/ctdchamacos). Incorporating study titles as part of CTD's exposure module is a simple but highly effective way to associate disparate publications, and users now can leverage this functionality as a new search parameter. Both of these new data fields are also included in the freely available, simplified view of the 'Exposure-study associations' data file (http://ctdbase.org/downloads/#exposurestudies).

Finally, we expanded the display content of 'Exposure Details' from 14 to 33 data fields and adapted the view mode to be customizable by the user (19). Initially, the four content areas of exposure curation (Stressor, Receptor, Event, and Outcome) displayed basic details such as the chemical stressor, description of receptors, biomarker term, measured levels, and disease or phenotype outcome. Though additional data fields were collected at the time of curation, these were not displayed on public CTD. Significant expansion of our database architecture, incorporation of new tables and modification of existing ones, as well as introduction of new validation processes, now enable public access to the full exposure-related dataset. The additional fields include details such as stressor source, receptor age, sex, race and smoking status, assay methods, limits of detection, and phenotype anatomy. New search features have also been incorporated into query pages to search 21 of the different 'Exposure Details' data fields. Currently, web pages automatically load with a default setting of 14 data columns (e.g. Mercury: http://ctdbase.org/detail.go?type=chem&acc=D008628&view=expConsol), but users can now specify which report fields are shown and accordingly adjust the display to include any set of the 33 available data categories. This allows investigators to more readily compare exposure measurements for specific parameters, such as enrollment years, race/ethnicity, detection frequency, U.S. state, etc. This expanded content of 'Exposure-event associations' is also freely available to download (http://ctdbase.org/downloads/#exposureevents).

Additional disease mappings in MEDIC

Since 2006, CTD has maintained and used MEDIC as a practical vocabulary for curating disease information (15). We created MEDIC by merging disease terms from the flat list of the OMIM resource (39) into the MeSH disease hierarchy (35) to produce an extensive, navigable controlled vocabulary. In the initial construction of MEDIC, we only mapped OMIM disease terms that had a known gene association. To help synchronize MEDIC with a more current version of OMIM, we routinely re-examine the OMIM source files to determine if additional diseases with associated genes had been added. In December 2017, CTD manually reviewed and mapped 1717 new OMIM terms into MEDIC following established protocols (15). MEDIC is still proving to be remarkably successful, convenient, robust, and adaptable, and we continue to create cross-references between MEDIC terms (MESH:ID and OMIM:ID) and the Disease Ontology (40) to help build a more comprehensive, interoperable (e.g. Figure 2C), and complete disease vocabulary for the scientific community. The improved MEDIC is freely available as a data file from CTD (http://ctdbase.org/downloads/#alldiseases).

Computational resources

The increased data content, coupled with the new and complex interrelationships between the core CTD datasets and the new phenotype- and exposure-related functionality, taxed CTD's existing computational infrastructure. In order accommodate the increased processing load, and to provide enhanced computational capacity moving forward, significant upgrades were made to the infrastructure. Older servers were replaced with current generation server blades that integrate more and faster CPUs with larger memory capacity. Additional storage capacity for the database was also procured in the form of new, faster storage arrays, and the server operating systems, PostgreSQL database servers, Tomcat application servers, and all 3rd party libraries were tested and upgraded to current versions.

FUTURE DIRECTIONS

CTD's first and foremost priority is always to increase data content, improve data completeness, and maintain data currency to keep CTD relevant, comprehensive, and up-to-date in an efficient manner.

Additionally, we will explore ways to make phenotype content even more functional and visible. One aim is to incorporate and display 'Inferred Phenotypes' as a new data-tab on CTD disease pages. This will tabulate phenotypes that can be inferred to diseases based upon shared chemicals (e.g. Figure 2D) or shared genes, and will allow users

to visualize genes, chemicals, phenotypes, and diseases in a single row, accommodating the construction of potential AOPs. As well, we plan to add phenotype content as input and output options to our suite of analytical and visualization tools, such as *Batch Query* and *VennViewer* (http://ctdbase.org/tools/).

Finally, we continually investigate ways in which to curate and produce additional content and facilitate exploration of our dataset. To wit, while manually curating the scientific literature, CTD biocurators often encounter unique author-created synonyms, abbreviations, or acronyms for chemicals, and many of these new phrases are not part of the current synonym list in CTD. Collecting and adding these new chemical terms to our chemical controlled vocabulary will not only help biocurators and users identify their chemical-of-interest in the database, but also enable us to construct a more comprehensive, relevant dictionary of chemical phrases for the scientific community and CTD text-mining projects (41). Toward that end, we plan to develop and implement a new 'chemical synonym collection' feature to the CTD Curation Tool (6), allowing biocurators to easily and rapidly collect newly encountered literature-based synonyms, abbreviations, acronyms, and CAS numbers, together with the associated PubMed article (for traceability). Ultimately, these new terms will be displayed on CTD official chemical pages and be searchable in all query forms. Other potential enhancements include further contextualizing CTD curated interactions with *'in vitro/in vivo'* status and adding taxon information to our collected disease annotations (to highlight animal models of disease). We have been collecting these data for a number of years, but have not yet made them public, and we are exploring the best ways to integrate this value-added content throughout the CTD framework.

SUMMARY

1. CTD curated content has increased by 46% and now provides more than 38 million toxicogenomic relationships.
2. CTD has launched a new chemical–phenotype module, with over 169 000 manually curated interactions relating 6700 chemicals, 4200 non-disease phenotypes, and 790 anatomical terms for 230 taxa; these data can be used both for knowledge discovery and generating predictive AOPs, and is searchable with an advanced query form.
3. CTD exposure content has increased by 74%, with additional functionality from two new querying parameters (*Study Factors* and *Associated Study Titles*) for 'Exposure Studies' and expanded content, query capability, and customizable web display options for 'Exposure Details'.
4. CTD's MEDIC disease controlled vocabulary has been updated with >1700 additional diseases and accession identifiers from OMIM.
5. CTD has upgraded its computational infrastructure to accommodate data content and functionality.

CITING AND LINKING TO CTD

To cite CTD data, please see: http://ctdbase.org/about/publications/#citing. If you are interested in establishing links to CTD data, please notify us (http://ctdbase.org/help/contact.go) and follow these instructions: http://ctdbase.org/help/linking.jsp.

FUNDING

National Institute of Environmental Health Sciences [R01 ES014065, R01 ES023788, P30 ES025128]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Funding for open access charge: National Institute of Environmental Health Sciences [R01 ES014065 and R01 ES023788].

Conflict of interest statement. None declared.

REFERENCES

1. Davis,A.P., Murphy,C.G., Saraceni-Richards,C.A., Rosenstein,M.C., Wiegers,T.C. and Mattingly,C.J. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic Acids Res.*, **37**, D786–D792.
2. Davis,A.P., King,B.L., Mockus,S., Murphy,C.G., Saraceni-Richards,C., Rosenstein,M., Wiegers,T. and Mattingly,C.J. (2011) The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.*, **39**, D1067–D1072.
3. Davis,A.P., Murphy,C.G., Johnson,R., Lay,J.M., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Rosenstein,M.C., Wiegers,T.C. *et al.* (2013) The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.*, **41**, D1104–D1114.
4. Davis,A.P., Grondin,C.J., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Wiegers,T.C. and Mattingly,C.J. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914–D920.
5. Davis,A.P., Grondin,C.J., Johnson,R.J., Sciaky,D., King,B.L., McMorran,R., Wiegers,J., Wiegers,T.C. and Mattingly,C.J. (2017) The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res.*, **45**, D972–D978.
6. Davis,A.P., Wiegers,T.C., Rosenstein,M.C., Murphy,C.G. and Mattingly,C.J. (2011) The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database*, **2011**, bar034.
7. Mattingly,C.J., Rosenstein,M.C., Davis,A.P., Colby,G.T., Forrest,J.N. and Boyer,J.L. (2006) The Comparative Toxicogenomics Database: a cross-species resource for building chemical–gene interaction networks. *Toxicol. Sci.*, **92**, 587–595.
8. Mattingly,C.J., Rosenstein,M.C., Colby,G.T., Forrest,J.N. and Boyer,J.L. (2006) The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J. Exp. Zool. A Comp. Exp. Biol.*, **305**, 689–692.
9. Davis,A.P., Murphy,C.G., Rosenstein,M.C., Wiegers,T.C. and Mattingly,C.J. (2008) The Comparative Toxicogenomics Database facilitates identification and understanding of chemical–gene–disease associations: arsenic as a case study. *BMC Med. Genomics*, **1**, 48.
10. Davis,A.P., Murphy,C.G., Saraceni-Richards,C.A., Rosenstein,M.C., Wiegers,T.C., Hampton,T.H. and Mattingly,C.J. (2009) GeneComps and ChemComps: a new CTD metric to identify genes and chemicals with shared toxicogenomic profiles. *Bioinformation*, **4**, 173–174.
11. Davis,A.P., Wiegers,T.C., Roberts,P.M., King,B.L., Lay,J.M., Lennon-Hopkins,K., Sciaky,D., Johnson,R., Keating,H., Greene,N. *et al.* (2013) A CTD-Pfizer collaboration: manual curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions. *Database*, **2013**, bat080.
12. Davis,A.P., Wiegers,T.C., Wiegers,J., Johnson,R.J., Sciaky,D., Grondin,C.J. and Mattingly,C.J. (2018) Chemical-induced phenotypes at CTD help to inform the pre-disease state and construct adverse outcome pathways. *Toxicol. Sci.*, **165**, 145–156.
13. Gohlke,J.M., Thomas,R., Zhang,Y., Rosenstein,M.C., Davis,A.P., Murphy,C., Becker,K.G., Mattingly,C.J. and Portier,C.J. (2009) Genetic and environmental pathways to complex diseases. *BMC Syst. Biol.*, **3**, 46.

14. Davis,A.P., Rosenstein,M.C., Wiegers,T.C. and Mattingly,C.J. (2011) DiseaseComps: a metric that discovers similar diseases based upon common toxicogenomic profiles at CTD. *Bioinformation*, **7**, 154–156.
15. Davis,A.P., Wiegers,T.C., Rosenstein,M.C. and Mattingly,C.J. (2012) MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database*, **2012**, bar065.
16. Davis,A.P., Wiegers,T.C., King,B.L., Wiegers,J., Grondin,C.J., Sciaky,D., Johnson,R.J. and Mattingly,C.J. (2016) Generating Gene Ontology-disease inferences to explore mechanisms of human disease at the Comparative Toxicogenomics Database. *PLoS One*, **11**, e0155530.
17. Mattingly,C.J., McKone,T.E., Callahan,M.A., Blake,J.A. and Hubal,E.A. (2012) Providing the missing link: the exposure science ontology ExO. *Environ. Sci. Technol.*, **46**, 3046–3053.
18. Grondin,C.J., Davis,A.P., Wiegers,T.C., King,B.L., Wiegers,J.A., Reif,D.M., Hoppin,J.A. and Mattingly,C.J. (2016) Advancing exposure science through chemical data curation and integration in the Comparative Toxicogenomics Database. *Environ. Health Perspect.*, **124**, 1592–1599.
19. Grondin,C.J., Davis,A.P., Wiegers,T.C., Wiegers,J.A. and Mattingly,C.J. (2018) Accessing an expanded exposure science module at the Comparative Toxicogenomics Database. *Environ. Health Perspect.*, **126**, 014501.
20. Planchart,A., Green,A., Hoyo,C. and Mattingly,C.J. (2018) Heavy metal exposure and metabolic syndrome: evidence from human and model system studies. *Curr. Environ. Health Rep.*, **5**, 110–124.
21. Pelletier,D., Wiegers,T.C., Enayetallah,A., Kibbey,C., Gosink,M., Koza-Taylor,P., Mattingly,C.J. and Lawton,M. (2016) ToxEvaluator: an integrated computational platform to aid the interpretation of toxicology study-related findings. *Database*, **2016**, baw062.
22. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
23. Mattingly,C.J., Boyles,R., Lawler,C.P., Haugen,A.C., Dearry,A. and Haendel,M. (2016) Laying a community-based foundation for data-driven semantic standards in environmental health sciences. *Environ. Health Perspect.*, **124**, 1136–1140.
24. McQuilton,P., Gonzalez-Beltran,A., Rocca-Serra,P., Thurston,M., Lister,A., Maguire,E. and Sansone,S.A. (2016) BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database*, **2016**, baw075.
25. Gaudet,P., Bairoch,A., Field,D., Sansone,S.A., Taylor,C., Attwood,T.K., Bateman,A., Blake,J.A., Bult,C.J., Cherry,J.M. *et al.* (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Database*, **2011**, baq027.
26. Davis,A.P., Johnson,R.J., Lennon-Hopkins,K., Sciaky,D., Rosenstein,M.C., Wiegers,T.C. and Mattingly,C.J. (2012) Targeted journal curation as a method to improve data currency at the Comparative Toxicogenomics Database. *Database*, **2012**, bas051.
27. Davis,A.P., Wiegers,T.C., Johnson,R.J., Lay,J.M., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., Murphy,C.G. and Mattingly,C.J. (2013) Text mining effectively scores and ranks the literature for improving chemical–gene–disease curation at the Comparative Toxicogenomics Database. *PLoS One*, **8**, e58201.
28. King,B.L., Davis,A.P., Rosenstein,M.C., Wiegers,T.C. and Mattingly,C.J. (2012) Ranking transitive chemical–disease inferences using local network topology in the Comparative Toxicogenomics Database. *PLoS One*, **7**, e46524.
29. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
30. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
31. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Kominger,F., May,B. *et al.* (2018) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
32. Chatr-Aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C., Sellam,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.
33. Galperin,M.Y., Fernandez-Suarez,X.M. and Rigden,D.J. (2017) The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. *Nucleic Acids Res.*, **45**, D1–D11.
34. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
35. Coletti,M.H. and Bleich,H.L. (2001) Medical subject headings used to search the biomedical literature. *J. Am. Med. Inform. Assoc.*, **8**, 317–323.
36. Agarwala,R., Barrett,T., Beck,J., Benson,D.A., Bollin,C., Bolton,E., Bourexis,D., Brister,J.R., Bryant,S.H., Canese,K. *et al.* (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **46**, D8–D13.
37. Deans,A.R., Lewis,S.E., Huala,E., Anzaldo,S.S., Ashburner,M., Balhoff,J.P., Blackburn,D.C., Blake,J.A., Burleigh,J.G., Chanet,B. *et al.* (2015) Finding our way through phenotypes. *PLoS Biol.*, **13**, e1002033.
38. Oki,N.O., Nelms,M.D., Bell,S.M., Mortensen,H.M. and Edwards,S.W. (2016) Accelerating adverse outcome pathway development using publicly available data sources. *Curr. Environ. Health Rep.*, **3**, 53–63.
39. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
40. Kibbe,W.A., Arze,C., Felix,V., Mitraka,E., Bolton,E., Fu,G., Mungall,C.J., Binder,J.X., Malone,J., Vasant,D. *et al.* (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, **43**, D1071–D1078.
41. Wiegers,T.C., Davis,A.P., Cohen,K.B., Hirschman,L. and Mattingly,C.J. (2009) Text mining and manual curation of chemical–gene–disease networks for the Comparative Toxicogenomics Database (CTD). *BMC Bioinformatics*, **10**, 326.