D972–D978 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 19 September 2016*
doi: 10.1093/nar/gkw838

The Comparative Toxicogenomics Database: update 2017

Allan Peter Davis[1,*], Cynthia J. Grondin[1], Robin J. Johnson[1], Daniela Sciaky[1], Benjamin L. King[2], 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 9, 2016; Accepted September 9, 2016

ABSTRACT

The Comparative Toxicogenomics Database (CTD; http://ctdbase.org/) provides information about interactions between chemicals and gene products, and their relationships to diseases. Core CTD content (chemical-gene, chemical-disease and gene-disease interactions manually curated from the literature) are integrated with each other as well as with select external datasets to generate expanded networks and predict novel associations. Today, core CTD includes more than 30.5 million toxicogenomic connections relating chemicals/drugs, genes/proteins, diseases, taxa, Gene Ontology (GO) annotations, pathways, and gene interaction modules. In this update, we report a 33% increase in our core data content since 2015, describe our new exposure module (that harmonizes exposure science information with core toxicogenomic data) and introduce a novel dataset of GO-disease inferences (that identify common molecular underpinnings for seemingly unrelated pathologies). These advancements centralize and contextualize real-world chemical exposures with molecular pathways to help scientists generate testable hypotheses in an effort to understand the etiology and mechanisms underlying environmentally influenced diseases.

INTRODUCTION

The Comparative Toxicogenomics Database (CTD; http://ctdbase.org) is a public resource for toxicogenomic information manually curated from the peer-reviewed scientific literature, providing key information about the interactions of environmental chemicals with gene products and their effect on human disease (1–4). CTD is curated by professional biocurators who leverage controlled vocabularies, ontologies and structured notation to code a triad of core interactions describing chemical-gene, chemical-disease and gene-disease relationships (5), which are then internally integrated to generate inferred chemical-gene-disease networks. These data are further associated with external data sets to establish novel, statistically ranked inferences between diverse types of information (6–7). Additionally, as part of our continued, active engagement with the scientific community, CTD plays a significant role in advancing text-mining methods for biomedical information as part of the BioCreative consortium (8–12), facilitates the development of semantic standards for the environmental health science community (13), complies with reporting standards set by the BioSharing Information Resources (14) and is a registered member (https://biosharing.org/biodbcore-000173) of BioDBcore (15).

Here, we provide our biennial database update, most notably highlighting our newly released exposure science module, which harmonizes and integrates data on chemical exposures into CTD's broader biological framework (16). Exposure science plays an important role in evaluating experimental toxicity data, developing risk assessments and informing public health policy (17). Centralization of human exposure information is critical to assessing the 'exposome', defined as the cumulative measure of an individual's exposure since birth (18). As well, the exposome complements genome research by recording and measuring the environmental component with which genes interact to determine one's phenotype (18). CTD's new exposure science module provides investigators with a centralized resource for making connections between real-world environmental chemical measurements and laboratory-derived toxicogenomic data. This new feature, as well as other updates described herein, further expands the utility of CTD for environmental health research.

*To whom correspondence should be addressed. Tel: +1 207 288 9880 (Ext. 128); Fax: +1 207 288 2130; Email: apdavis3@ncsu.edu
Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

NEW FEATURES

Increased data content, dissemination and use by scientific community

In July 2016, core CTD consisted of over 1.6 million manually curated interactions (including 1 379 105 chemical-gene, 202 085 chemical-disease and 33 583 gene-disease direct interactions) for 14 672 chemicals, 42 761 genes and 6401 diseases coded from 117 866 peer-reviewed scientific articles studied in 564 species (Table 1). This represents a 33% increase in chemical-gene-disease interactions since our last update (4). CTD also integrates manually curated data to generate predictive inferences (6,7); for example, if chemical A interacts with gene B, and independently gene B is associated with disease C, then chemical A is inferred to have a relationship with disease C (via gene B). Internal integration of core data generated more than 19.7 million inferred gene-disease relationships and 1.8 million inferred chemical-disease relationships, which are statistically ranked (7). Finally, integration with external annotations from GO (19), KEGG (20), Reactome (21) and BioGRID (22) yields additional inferred relationships (Table 1). In total, more than 30.5 million toxicogenomic connections are now freely available for analysis and hypothesis development.

Aside from our own public web application (PWA), CTD research and curated content is disseminated further into the scientific community in a number of important ways. First, at least 72 external resources now include and display CTD information as part of their own databases, a 44% increase from the 50 sources since our last report (http://ctdbase.org/about/publications/#use). Second, in conjunction with Pfizer scientists, we developed ToxEvaluator, a proprietary tool that integrates CTD chemical-gene-disease relationships with other diverse (public and private) datasets into a single, web-based platform to aid Pfizer scientists in generating mechanistic toxicity-related hypotheses (23). Finally, CTD continues its commitment to spearhead and advance biomedical text-mining research for the scientific community by teaming with the National Center for Biotechnology Information (NCBI) to organize a BioCreative community challenge which focused on developing tools to identify and extract specific disease and chemical content (11). For this endeavor, we helped develop a large corpus of manually curated annotations for chemicals, diseases and their interactions from 1500 PubMed articles (12); this corpus is freely available (download by clicking: http://sourceforge.net/projects/bioc/files/CDR_Data.zip/download), as are many of the associated text-mining tools developed by the 25 teams that participated.

New CTD exposure science module

Most notably, since our last update, CTD has released a new exposure module (16). This component was developed in response to the community's need for a centralized database that curates and harmonizes the real-world measurements and biological effects of environmental chemicals (e.g. air pollutants, pesticides, heavy metals, polychlorinated biphenyls, *inter alia*) and human biomarkers incurred by such exposure.

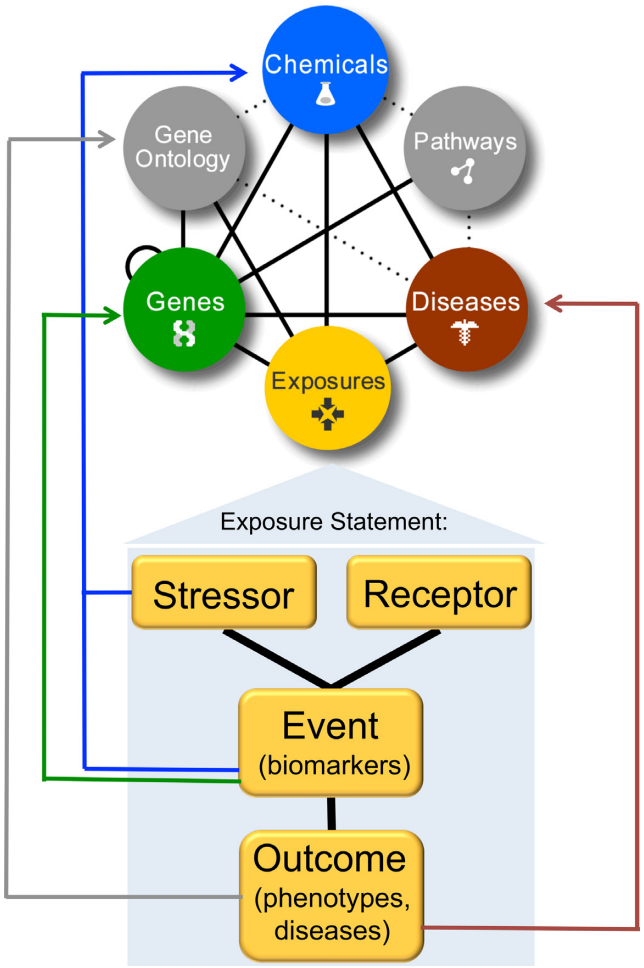

Figure 1. Integration of exposure data curation into CTD framework. Core CTD is composed of interactions between chemicals, genes, diseases, Gene Ontology (GO) terms and pathway annotations (colored circles in top diagram). In March 2015, CTD released an exposure science module (light orange circle in top diagram). For this paradigm, CTD manually curates exposure statements (bottom light blue box, with light orange categories connected by black lines) describing how environmental stressors interact with human receptors during an exposure event to result in an exposure outcome. This curation paradigm uses many of the same controlled vocabularies as those used in core CTD to allow seamless data integration and connectivity between the two projects: exposure stressors are chemicals (blue arrow); exposure events report biomarker measurements for chemicals (blue arrow) and proteins (green arrow); and exposure outcomes can be either altered phenotypes (defined as GO terms, gray arrow) or diseases (red arrow).

Working with the exposure science community, CTD developed a novel manual curation paradigm (16) using the exposure ontology as its foundation (24). In this module, CTD biocurators annotate over 35 data fields to four main categories that collectively form an exposure statement. A statement relates how an exposure stressor interacts with a human exposure receptor during an exposure event to result in an exposure outcome (Figure 1). An integral feature of this curation paradigm is that we use many of the same controlled vocabularies when curating chemical-gene-disease interactions for core CTD (5,16). Thus, chem-

Table 1. Updated core CTD content (July 2016)

Source	Data type	Counts
Manual curation	Scientific articles	117 866
Manual curation	Chemicals	14 672
Manual curation	Genes	42 761
Manual curation	Diseases	6401
Manual curation	Taxa	561
Manual curation	Chemical-gene interactions	1 379 105
Manual curation	Gene-disease interactions	33 583
Manual curation	Chemical-disease interactions	202 085
Data integration	Gene-disease inferences	19 720 041
Data integration	Chemical-disease inferences	1 858 286
Data integration	Chemical-GO inferences	4 529 027
Data integration	Chemical-pathway inferences	307 728
Data integration	Disease-pathway inferences	59 863
Data integration	Disease-GO inferences	795 845
Imported	Gene-GO annotations	1 201 527
Imported	Gene-pathway annotations	63 863
Imported	Gene-gene interactions	376 472
Total		30 527 425

icals described as an exposure stressor or event biomarker are annotated with CTD's chemical vocabulary; similarly, protein biomarkers (e.g. serum proteins, cytokines, interleukins) are coded using CTD's gene vocabulary; and exposure outcomes are annotated to either CTD's MEDIC disease vocabulary (25) or GO biological process (GO-BP) terms for phenotypes, which we have previously defined as 'non-disease-term biological events' (26). This curation strategy provides three important advantages. First, it allows heterogeneous exposure information from different articles published by different laboratories in different journals over the decades to become standardized and centralized into a single repository, facilitating connections between unique studies. Second, it brings exposure science data into the broader CTD framework, allowing both exposure data to leverage CTD curated knowledge and also allowing core CTD to help inform exposure analysis. Finally, the use of controlled vocabularies transforms complex, interdependent exposure incidents into modular data, allowing exposure information to be sorted, filtered and viewed from a variety of perspectives (such as geographical location and receptor type).

The manually curated exposure data are displayed on two new tabs on CTD's PWA: 'Exposure Studies' (providing a summary of each exposure article) and 'Exposure Details' (providing detailed biomarker measurements), on all relevant Chemical, Gene, Disease, GO and Reference pages. Additionally, investigators can use CTD's new exposure study query page (http://ctdbase.org/query.go?type=expStudies) to quickly retrieve information aggregated at the study (research article) level (Figure 2A) using parameters for a chemical stressor (e.g. 'air pollutants'), the type of human receptor studied (e.g. 'study subjects') and a geographic location (e.g. 'United States'). Select terms (chemicals, genes, diseases, GO and references) returned in the result page are hyperlinked to their corresponding pages in CTD (Figure 2B), allowing users to seamlessly explore additional associated information.

The real-world measurements of exposure biomarkers are found under the 'Details' link on an 'Exposure Studies' page or can be viewed *in toto* on the 'Exposure Details' data-tab for a chemical-of-interest (Figure 3). This latter option aggregates the data from the germane articles in CTD, providing users with a landscape view of exposure measurements and outcomes from the published literature. As well, an 'Exposure Details' query page is also available (http://ctdbase.org/query.go?type=expDetails) to retrieve records at the highly granular exposure statement level.

In July 2016, CTD included more than 70 600 manually curated exposure statements for 803 chemicals, 153 biomarker genes, 301 diseases and 181 phenotypes (GO-BP terms) from over 1250 exposure scientific articles (Figure 2C). Exposure data is also freely available to users in download files (http://ctdbase.org/downloads/#exposureevents).

New availability of GO-Disease inferences

Over the last decade (4), CTD has successfully generated novel connections between different types of data by integrating diverse information through a common intermediate (1,6,7). For example, if gene A is annotated to GO biological process term B (by GO annotators), and, independently, gene A is also associated with disease C (by CTD biocurators), then GO term B can be inferred to disease C (via gene A) (Figure 4A). These GO-disease inferences help users discover common molecular, biological and cellular events shared among seemingly unrelated diseases (Figure 4B). The availability of this novel dataset (27) can be leveraged in numerous ways, including discovering potential comorbidities (especially in exposure science), possible new treatment options by repositioning pharmaceutical drugs or identifying possible side effects. CTD's files for GO-Disease-Gene Inference Networks are freely available (http://ctdbase.org/downloads/#godiseasegenes) and in July 2016 included more than 795 000 inferences between over 15 700 GO terms and 4200 diseases.

Disease mappings and link-outs

Since 2006, CTD has maintained and used MEDIC as a practical vocabulary for curation of disease information (25). MEDIC was created by merging disease terms from

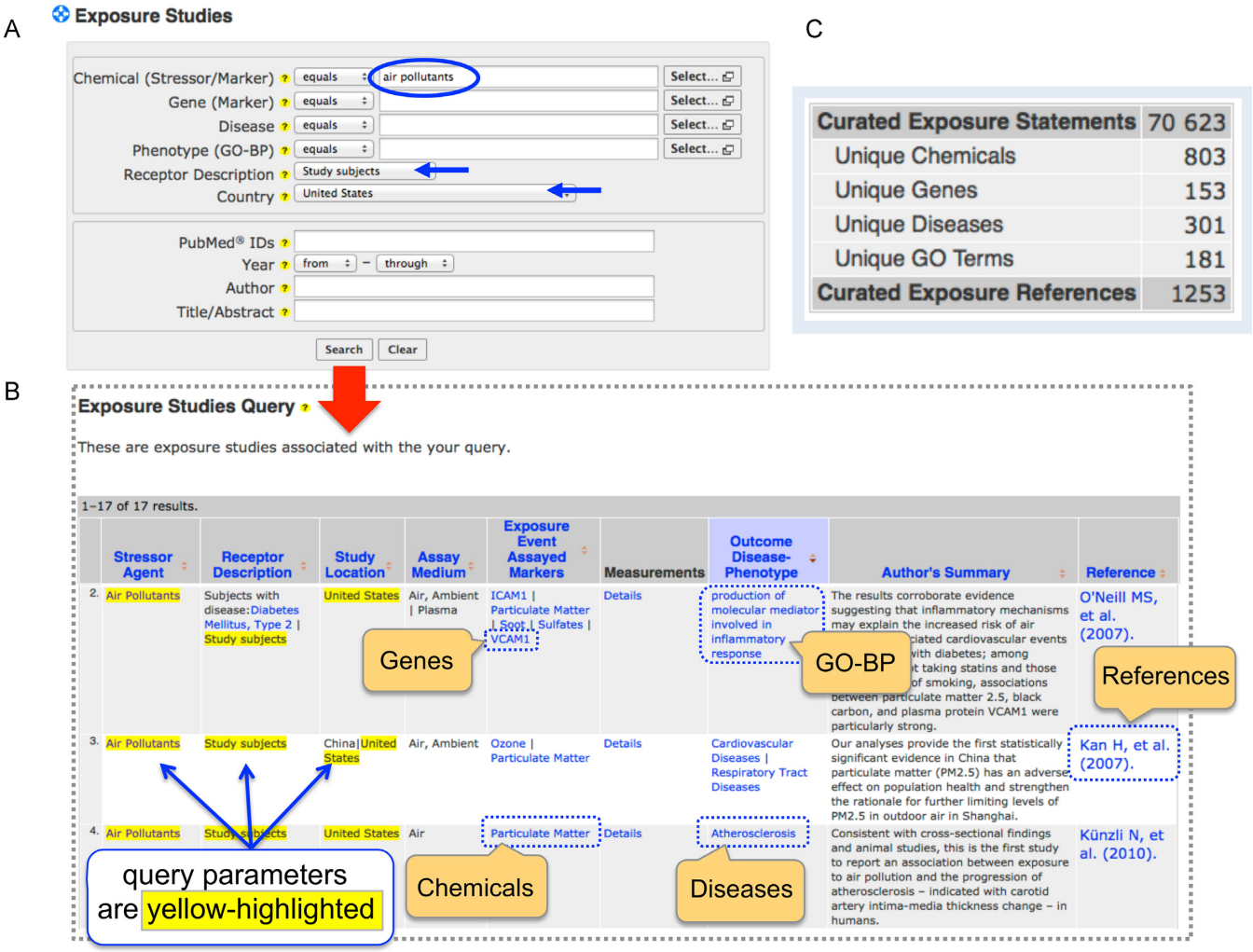

Figure 2. CTD's new exposure curation content. (**A**) CTD's new 'Exposure Studies' query page (http://ctdbase.org/query.go?type=expStudies) allows users to retrieve study information using a variety of search parameters, including exposure chemical stressor (e.g. air pollutants, blue circle), receptor description (e.g. study subjects) and countries (e.g. the USA) as the study location (blue arrows). (**B**) The results display curated exposure studies meeting the query parameters (highlighted in yellow) with integrated links to any mentioned genes, chemicals, diseases, phenotypes (GO-BP terms) and source references (blue dotted boxes with orange callout labels). The 'Author's Summary' column provides the take-home point of each study, and the 'Details' link in the 'Measurements' column takes the user to the specific assay measurements. As well, all exposure content is reciprocally displayed on the aforementioned pages (i.e. gene, chemical, disease, phenotype and source references), making the information seamlessly integrated into the broader biological context of the CTD framework. (**C**) Data status for CTD exposure module as of July 2016 includes over 70 600 manually curated statements from more than 1250 references (updated monthly at: http://ctdbase.org/about/dataStatus.go).

the flat list of the OMIM resource (28) with two Medical Subject Heading (MeSH) disease hierarchies (29) to produce an extensive, navigable vocabulary. While originally intended to be only a placeholder until a more sophisticated disease resource emerged, MEDIC has proven to be remarkably successful, convenient and adaptable, and has been incorporated by many systems (30–34). In 2015, CTD began analyzing and comparing the disease terms and hierarchical structure used in MEDIC against the newly established Disease Ontology (DO) (35), in an effort to coordinate MEDIC with this new resource. A single, robust, community-accepted disease vocabulary would be valuable for synchronizing the vast arrays of different biological databases. Toward that end, CTD is coordinating with the DO staff to find ways that MEDIC could help inform DO, and vice versa. As a first step, CTD now provides direct web links between 3,258 MEDIC disease terms to 2,943 equivalent terms in DO, based upon common MeSH accession identification numbers shared between the two vocabularies. Ultimately, bidirectional cross-links between MEDIC and DO will enable greater interoperability and data sharing for the entire scientific community.

FUTURE DIRECTIONS

Since 2004, CTD has evolved from a fledgling database to an extensive public resource with over 30.5 million toxicogenomic relationships (Table 1). We will continue to expand our core and exposure curation modules with new data content added every month (http://ctdbase.org/about/dataStatus.go).

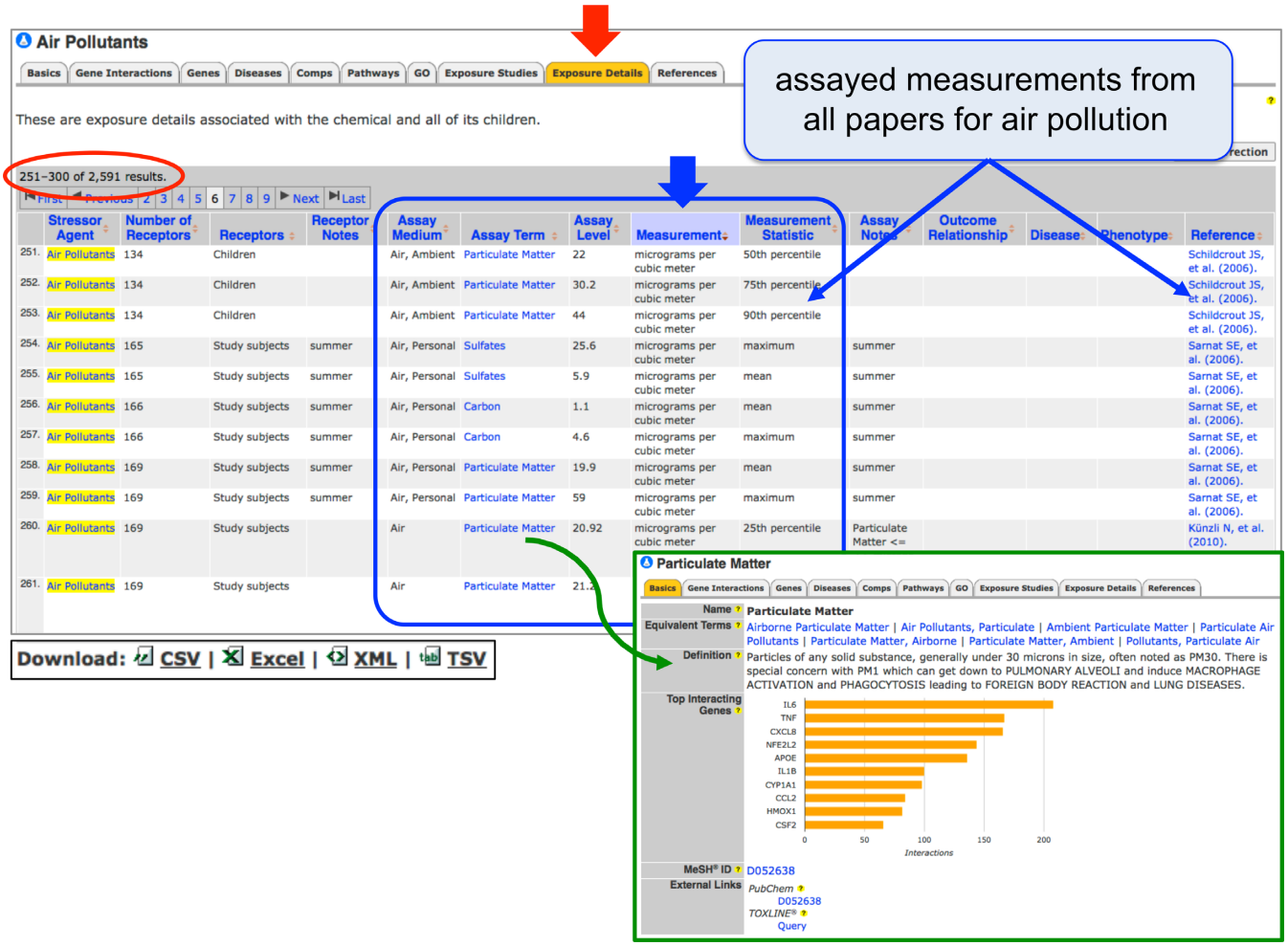

Figure 3. Landscape view of real-world measurements and outcomes curated for exposure science. The 'Exposure Details' data-tab (red arrow) on CTD's chemical page for air pollutants lists 2591 results (red circle) for air pollution markers (e.g. particulate matter, sulfates and carbon), including the type of medium in which the marker was assayed, the units of measurements, the statistics associated with the measurement (blue box), as well as any disease/phenotype outcomes. This view aggregates the data from all germane articles for the chemical-of-interest. Column headers in the table will sort the information by clicking (blue arrow) and embedded terms are hyperlinked to their respective CTD pages (green arrow and green inset box), allowing users to easily navigate to other concepts. At the bottom of every CTD page, a link allows users to download the information onto a desktop in a variety of formats.

Additionally, we plan to enhance and develop new visualization and analytical tools to help users better explore our curated exposure science data. Two goals include allowing users to choose which data fields are displayed on a web page and enriching query pages to allow greater specification and filtering of returned data. We also plan to leverage web-based maps to view exposure chemicals, events and outcomes from a geographical perspective. Currently in CTD, we have exposure data for 109 countries and all 50 US states.

As well, we intend to release a new phenotype module, which will include our manual curation of chemicals regulating biological, cellular and physiological events in conjunction with anatomical descriptors. This feature will help associate and identify chemical-induced phenotypes that precede the clinical manifestation of a disease. We previously released an initial deposit of this dataset using MeSH terms as our phenotype descriptions (26); however, since then, we have mapped those terms to the more versatile GO-BP controlled vocabulary to reflect greater granularity and broader coverage of biological concepts.

Finally, we plan to devise computational programs that will systematically connect the spectrum of CTD curated content by linking chemical-gene initiating events, chemical-phenotype and gene-GO key events, chemical-disease events and exposure-level outcomes for populations. Such computationally predicted adverse outcome pathways (cpAOP) have been recently described for fatty liver disease using rat data (36). We hope to systematically expand upon this effort by leveraging CTD data to generate cpAOPs connecting chemicals to disease outcomes.

SUMMARY

- We increased CTD content by 33% to over 30.5 million toxicogenomic relationships.

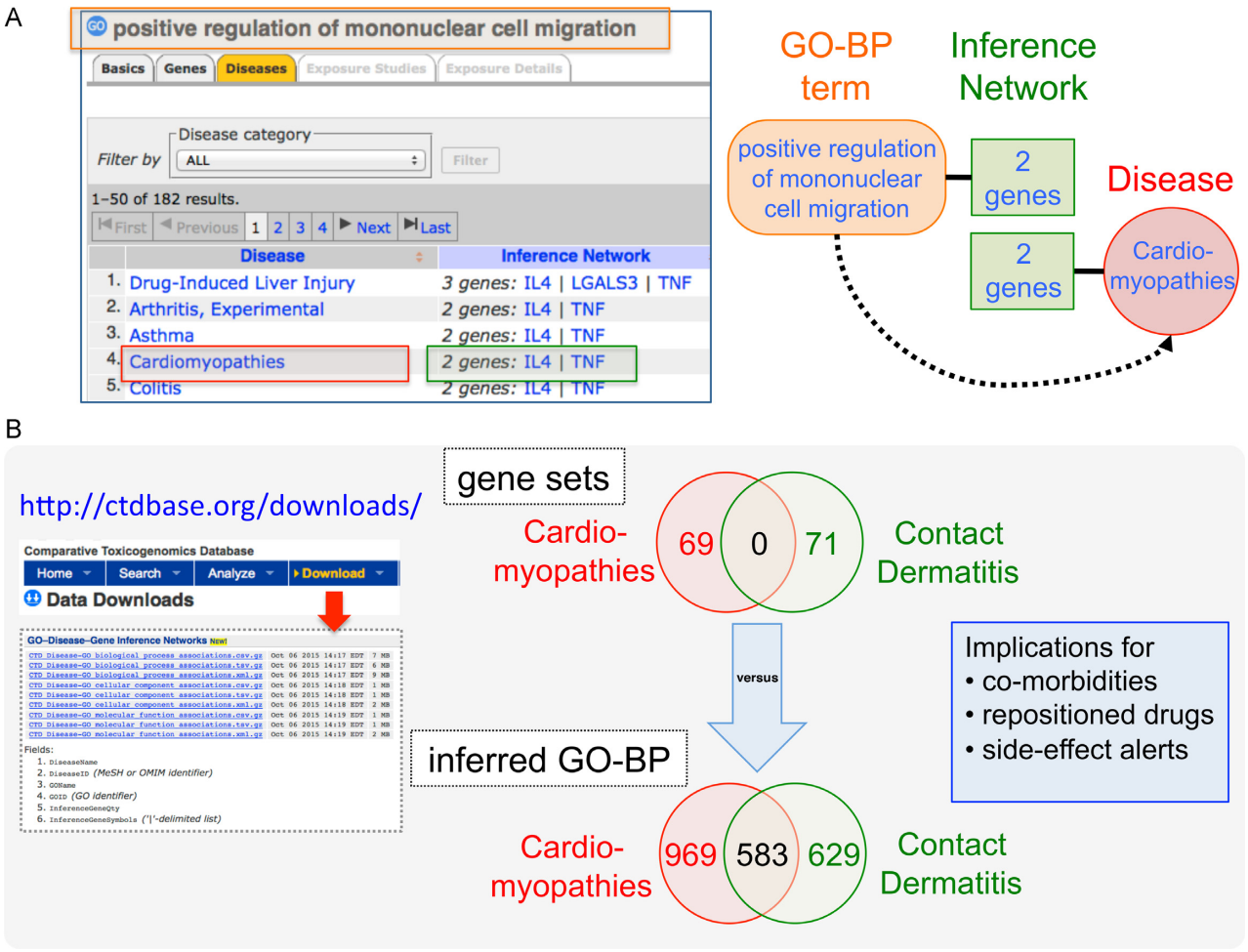

Figure 4. CTD's new dataset generates novel inferences between GO terms and diseases. (**A**) Every GO term has its own page in CTD, and GO terms can be inferred to diseases based upon shared genes. Data integration via shared genes (here, genes IL4 and TNF, green box) allows the GO biological process (GO-BP) term 'positive regulation of mononuclear cell migration' (orange oval) to be inferred to the disease cardiomyopathies (red circle). (**B**) An example of how users can leverage this information. Two diseases (cardiomyopathies and contact dermatitis) initially appear to be unrelated because they share no genes; however, when instead viewed using inferred GO-BP terms, the two diseases overlap significantly with 583 inferred GO-BP terms ($P = 4.72 \times 10^{-201}$, Fisher's exact test), suggesting potential molecular underpinnings common between the two pathologies. This discovery can have implications for recognizing co-morbidities (especially for exposure science), identifying avenues to reposition therapeutic drugs or creating alerts to potentially new side effects. CTD's files for 'GO-Disease-Gene Inference Networks' (as well as all CTD curated content) are freely available from our 'Data Downloads' page (http://ctdbase.org/downloads/).

- We introduced our exposure science module, containing 70 600 exposure statements for over 800 chemicals, 150 genes, 300 diseases and 180 phenotypes.
- We described our GO-disease inference dataset, connecting functional, biological and cellular events between seemingly unrelated diseases.
- We enhanced community database interoperability by providing links from our MEDIC disease vocabulary to DO terms.

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 ES019604, R01 ES023788]; National Institute of General Medical Sciences of the National Institutes of Health Institutional Development Awards [P20 GM103423, P20 GM104318 to B.L.K]; Funding for open access charge: National Institute of Environmental Health Sciences [R01 ES014065, R01 ES019604, 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. and Wiegers,T.C. et.al. (2013) The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.*, **41**, D1104–D1114.
4. Davis,A.P., Grondin,C.G., 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., 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*, doi:10.1093/database/bar034.
6. 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**, doi:10.1186/1755-8794-1-48.
7. 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.
8. Wiegers,T.C., Davis,A.P. and Mattingly,C.J. (2012) Collaborative biocuration-text-mining development task for document prioritization for curation. *Database*, doi:10.1093/database/bas037.
9. Arighi,C.N., Wu,C.H., Cohen,K.B., Hirschman,L., Krallinger,M., Valencia,A., Ju,Z., Wilbur,J.W. and Wiegers,T.C. (2014) BioCreative-IV virtual issue. *Database*, doi:10.1093/database/bau039.
10. Wiegers,T.C., Davis,A.P. and Mattingly,C.J. (2014) Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. *Database*, doi:10.1093/database/bau050.
11. Wei,C.H., Peng,Y., Leaman,R., Davis,A.P., Mattingly,C.J., Li,J., Wiegers,T.C. and Lu,Z. (2016) Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database*, doi:10.1093/database/baw032 .
12. Li,J., Sun,Y., Johnson,R.J., Sciaky,D., Wei,C.H., Leaman,R., Davis,A.P., Mattingly,C.J., Wiegers,T.C., Lu,Z. *et al.* (2016) BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, doi:10.1093/database/baw068.
13. 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.
14. 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*, doi:10.1093/database/baw075.
15. 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*, doi:10.1093/database/baq027.
16. 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.*, doi:10.1289/ehp174.
17. Hubal,E.A. (2009) Biologically relevant exposure science for 21st century toxicity testing. *Toxicol. Sci.*, **111**, 226–232.
18. Wild,C.P. (2005) Complementing the genome with an 'exposome': the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol. Biomarkers Prev.*, **14**, 1847–1850.
19. 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.
20. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,J. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
21. Fabregat,A., Sidiropoulos,K., Garapati,P., Gissespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Kominger,F., McKay,S. *et al.* (2016) The Reactome pathway Knowledgebase. *Nucleic Acids Res.*, **44**, D481–D462.
22. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
23. 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*, doi:10.1093/database/baw062.
24. 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.
25. 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*, doi:10.1093/database/bar065.
26. 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*, doi:10.1093/database/bat080.
27. 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.
28. 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.
29. 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.
30. Hayman,G.T., Laulederkind,S.J., Smith,J.R., Wang,S.J., Petri,V., Nigam,R., Tutaj,M., De Pons,J., Dwinell,M.R. and Shimoyama,M. (2016) The Disease Portals, disease-gene annotation and the RGD disease ontology at the Rat Genome Database. *Database*, doi:10.1093/database/baw034.
31. Leaman,R., Islamaj-Dogan,R. and Lu,Z. (2013) DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, **29**, 2909–2917.
32. Lee,H.C., Hsu,Y.Y. and Kao,H.Y. (2016) AuDis: an automatic CRF-enhanced disease normalization in biomedical text. *Database*, doi:10.1093/database/baw091.
33. Shimoyama,M., Smith,J.R., De Pons,J., Tutaj,M., Khampang,P., Hong,W., Erbe,C.B., Ehrlich,G.D., Bakaletz,L.O. and Kerschner,J.E. (2016) The Chinchilla Research Resource Database: resource for an otolaryngology disease model. *Database*, doi:10.1093/database/baw073.
34. Dai,H.J., Wu,J.C., Lin,W.S., Reyes,A.J., Dela Rosa,M.A., Syed-Abdul,S., Tsai,R.T. and Hsu,W.L. (2014) LiverCancerMarkerRIF: a liver cancer biomarker interactive curation system combining text mining and expert annotations. *Database*, doi:10.1093/database/bau085.
35. 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.
36. Bell,S.M., Angrish,M.M., Wood,C.E. and Edwards,S.W. (2016) Integrating publicly available data to generate computationally predicted adverse outcome pathways for fatty liver. *Toxicol. Sci.*, **150**, 510–520.