*Published online 22 September 2010* *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D1067–D1072***
*doi:10.1093/nar/gkq813*

# The Comparative Toxicogenomics Database: update 2011

**Allan Peter Davis, Benjamin L. King, Susan Mockus, Cynthia G. Murphy, Cynthia Saraceni-Richards, Michael Rosenstein, Thomas Wiegers and Carolyn J. Mattingly***

Department of Bioinformatics, The Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672, USA

Received July 29, 2010; Revised August 19, 2010; Accepted August 24, 2010

## ABSTRACT

**The Comparative Toxicogenomics Database (CTD) is a public resource that promotes understanding about the interaction of environmental chemicals with gene products, and their effects on human health. Biocurators at CTD manually curate a triad of chemical–gene, chemical–disease and gene–disease relationships from the literature. These core data are then integrated to construct chemical–gene–disease networks and to predict many novel relationships using different types of associated data. Since 2009, we dramatically increased the content of CTD to 1.4 million chemical–gene–disease data points and added many features, statistical analyses and analytical tools, including GeneComps and ChemComps (to find comparable genes and chemicals that share toxicogenomic profiles), enriched Gene Ontology terms associated with chemicals, statistically ranked chemical–disease inferences, Venn diagram tools to discover overlapping and unique attributes of any set of chemicals, genes or disease, and enhanced gene pathway data content, among other features. Together, this wealth of expanded chemical–gene–disease data continues to help users generate testable hypotheses about the molecular mechanisms of environmental diseases. CTD is freely available at http://ctd.mdibl.org.**

## INTRODUCTION

The environment is believed to play an important role in the etiology of many human diseases, and chemicals are an important component of the environment (1). The Comparative Toxicogenomics Database (CTD; http://ctd.mdibl.org) is a unique resource that makes connections between chemicals, gene products and diseases that may not otherwise be apparent, and provides the basis for testable hypotheses about the mechanisms underlying the etiology of environmental diseases (2–4).

Several valuable chemical databases currently exist (5). CTD is distinct in three important ways: it focuses on environmental chemicals, it manually curates and then integrates datasets to discover novel connections, and it functions as both a data repository as well as a tool for generating hypotheses about chemical actions and environmental diseases. The value and utility of CTD is evidenced by it being indexed at numerous other databases, including PubChem (6), PharmGKB (7), UniProt (8), T3DB (9), GAD (10) and ChemID (11), and by the inclusion of CTD's curated information in other database products, such as STITCH (12), ToppGene (13), PhenoHM (14), Chem2Bio2RDF (15), UCSC Browser (16), WhichGenes (17), ChemSpider (http://www.chemspider.com) and RefGene (http://refgene.com). As well, CTD datasets have been used by several independent groups for meta-analyses to derive relationships between environmental chemicals and complex human diseases (18–20). Furthermore, CTD will be included as a search option in the suite of integral databases at TOXNET, the National Library of Medicine's portal for toxicology data (21).

We previously reported on CTD in an introductory article (22); here we update the increased data content and describe several new analytical and visualization tools and enhancements to CTD since 2009.

## NEW FEATURES

### More curated data

The strength of CTD data still comes from information being derived by professional biocurators who read and

*To whom correspondence should be addressed. Tel: 207 288 3605; Fax: 207 288 2130; Email: cmattin@mdibl.org

**Table 1.** Increase in CTD data from 2008 to 2010

| Curated data | July 2010 | July 2008 |
|---|---|---|
| Publications | 21 668 | 8987 |
| Chemicals | 5931 | 3971 |
| Genes | 17321 | 13 300 |
| Diseases | 3818 | 2899 |
| Chemical–gene interactions | 240 369 | 116 067 |
| Direct gene–disease interactions | 11 573 | 5925 |
| Inferred gene–disease interactions | 886 604 | 349 300 |
| Direct chemical–disease interactions | 8523 | 2569 |
| Inferred chemical–disease interactions | 246 603 | 76 970 |
| Total chemical–gene–disease interactions | 1393 672 | 550 831 |

manually curate the peer-reviewed scientific literature. This process, albeit time-consuming, ensures that the core (triad) data, which underlie the establishment of novel relationships, are valid and accurate. Our curation paradigm, including methods and sources of controlled vocabularies and ontologies, and our integration strategy used to generate inferred relationships between datasets were described previously in detail (22). To streamline the curation process, CTD recently designed a text-mining application that efficiently prioritizes the vast toxicology literature (23), yet information is still extracted manually from articles. CTD curated data are then combined and integrated with other CTD data as well as annotations from external sources to produce a plethora of novel inferred relationships.

In July 2010, CTD contained over 240 300 molecular interactions between 5900 unique chemicals and 17 300 gene products, 11 500 direct gene–disease relationships and 8500 direct chemical–disease relationships extracted from over 21 600 publications (Table 1). Integration of these data generates 886 600 inferred gene–disease relationships and 246 600 inferred chemical–disease relationships. In total, ~1.4 million chemical–gene–disease data connections are now available for exploration and analysis, representing a 2.5-fold increase in the content since our original description.

## New data sources and links

In addition to the increased content, CTD has also expanded its external data sources. We now include pathway data from Reactome (24), in addition to the previously included KEGG pathways (25). Reactome and KEGG databases annotate genes into pathways, and this information is then integrated with CTD genes. By integrating these chemical–gene data with gene pathway data, novel chemical pathway connections are generated, allowing users to explore pathways that may be influenced by environmental chemicals. Additionally, CTD now includes links to gene pages at PharmGKB (7) and chemical pages at DrugBank (26).

## Enhanced data features

We have enhanced CTD data by adding three new computational features.

(i) *GeneComps and ChemComps*. Every curated gene and chemical now includes GeneComps and ChemComps data tabs, respectively. These new metrics statistically identify genes and chemicals with shared toxicogenomic profiles, allowing users to find genes and chemicals similar to their favorite molecule of interest as well as build toxicogenomic networks based upon shared molecular profiles (27). For example, the ubiquitous plastics component bisphenol A interacts with over 475 genes in CTD. ChemComps identifies other chemicals with similar gene interaction profiles and ranks them based upon their similarity index to produce a list of comparable chemicals that include polychlorinated biphenyls, genistein, estradiol and nonylphenol (27), supporting the known estrogenic behavior of bisphenol A.

(ii) *Enriched GO terms*. The gene ontology (GO) is a well-documented and widely used annotation system that assigns molecular function, biological process and cellular component information to gene products (28). Typically, GO annotations are used to retrieve and organize extensive biological knowledge about gene lists. Uniquely, we use GO data to understand the actions of non-genetic molecules (chemicals) by associating GO terms to chemicals via their curated interacting genes. The GO data-tab on chemical pages allows users to explore, sort and rank the enriched GO terms to gain a greater understanding of the biological effects of the chemical. For example, diazinon has over 500 enriched GO terms transferred to it via the 215 genes that interact with this insecticide (Figure 1). Only statistically significant GO terms are displayed with their calculated enrichment score (presented as the log-transformed probability from the hypergeometric distribution). Some of the deeper level GO terms include annotations describing dopamine metabolism and regulation (not shown), providing a link between diazinon and nerve cell processes that could help users generate testable hypotheses about insecticide exposure and neurological defects such as Parkinson's or Alzheimer's disease. In July 2010, CTD connected over 5000 enriched GO terms to more than 4500 chemicals.

(iii) *Inference network scores*. If chemical A has a curated interaction with gene B, and independently gene B is directly associated with disease C, then CTD integration generates an inferred relationship between chemical A and disease C (inferred via gene B). We now utilize local network topology-based statistics to evaluate these inferred chemical–disease relationships (King *et al*. in preparation and see Figure 2). The scores allow users to sort and rank the predicted chemical–disease relationships to help prioritize hypothesis testing. For example, tetrachlorodibenzodioxin (TCDD) has an inferred relationship with prostate cancer based on 65 commonly interacting genes (Figure 2). Notably, this high-ranking, inferred chemical–disease

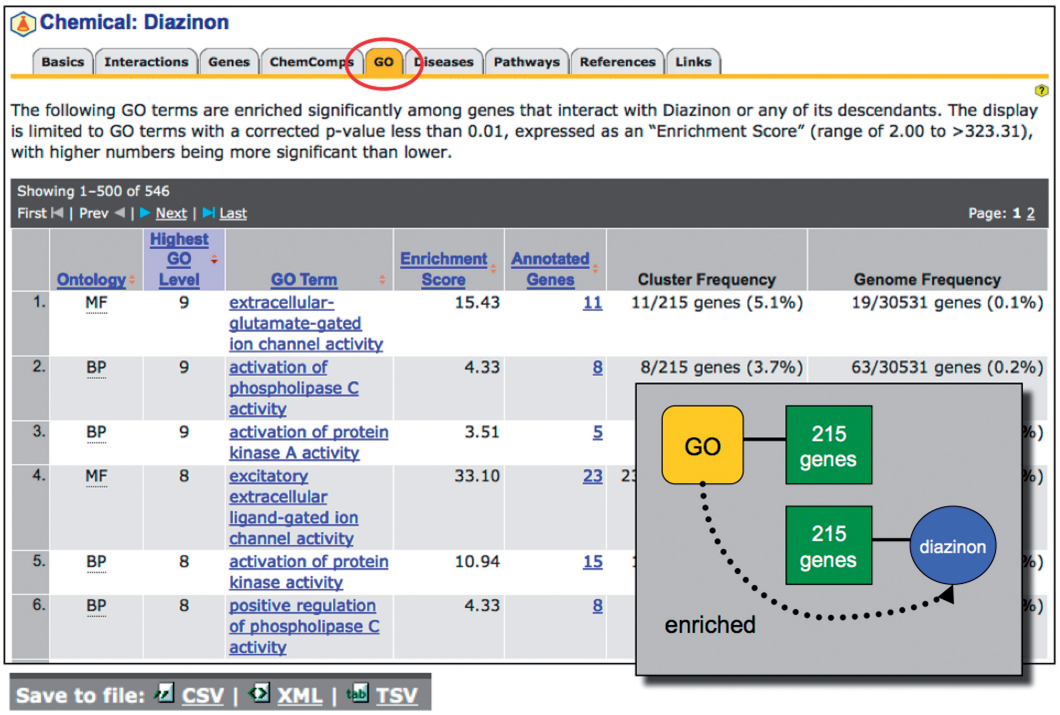


**Figure 1.** CTD's unique use of enriched GO terms for chemicals. On the highlighted 'GO' data tab (red circle) for a chemical, users will find enriched GO terms. Here, the insecticide diazinon has curated interactions with 215 genes. The GO annotations for these genes are statistically enriched and only the significant GO terms (*P*-value less than 0.01, expressed as an enrichment score) are shown for the chemical (inset). Cluster and genome frequency calculations are also provided. Users can sort the table by the column headers: Ontology, GO level, GO term, Enrichment score and Annotated genes. All CTD pages can be downloaded by using the 'Save to file' feature at the bottom of every page.

prediction was supported by a recent study correlating exposure to Agent Orange (which was contaminated with TCDD) and prostate cancer in US war veterans (29). In addition to providing potentially valid and novel disease associations, the genes that form the basis of chemical–disease inferences in CTD may help to explain the mechanisms underlying the etiology of a particular disease and provide a starting point to develop testable hypotheses about chemical effects on human health.

## New tools

To help navigate the 1.4 million chemical–gene–disease data points in CTD, we have created a suite of analytical and visualization tools, accessible from the 'Tools' menu bar.

(i) *Batch Query* has been previously described (22) and still represents an efficient way to download any type of data associated with a user's input list of chemicals, genes or diseases.

(ii) *VennViewer* is a new tool that allows users to create Venn diagrams to compare associated datasets for chemicals, genes or diseases. A user can input any three chemicals and determine the overlapping and unique sets of associated data for interacting genes, diseases, inferred KEGG and Reactome pathways, or enriched GO terms (Figure 3a). Similar comparisons can be made for lists of genes or diseases. The high-resolution images of the Venn diagram can be downloaded.

(iii) *MyGeneVenn* is similar to VennViewer; however, it allows users to compare their own gene list to gene sets in CTD that are associated with specified chemicals or diseases. For example, a user may have a set of genes from a microarray experiment that are regulated in response to bisphenol A. MyGeneVenn can be used to determine which genes from that set are known to interact with resveratrol or genistein (Figure 3b) or which genes have been previously shown to interact with bisphenol A. Similarly, MyGeneVenn will let the user resolve which genes from their set are associated with specified diseases.

## Other new features

CTD has been enhanced with many other features to make the website even more user-friendly, including a redesigned homepage to make navigation easier and more intuitive, a 'Downloads' menu tab that allows users to download all of CTD's dataset files, and a 'Help' menu tab with links to our FAQ (which includes step-by-step instructions on how to perform various queries), tutorial resources (including our handy Resource Guide for quick reference to CTD), instructions on how to link to CTD and a site to join our email list to stay informed about new features and releases, including

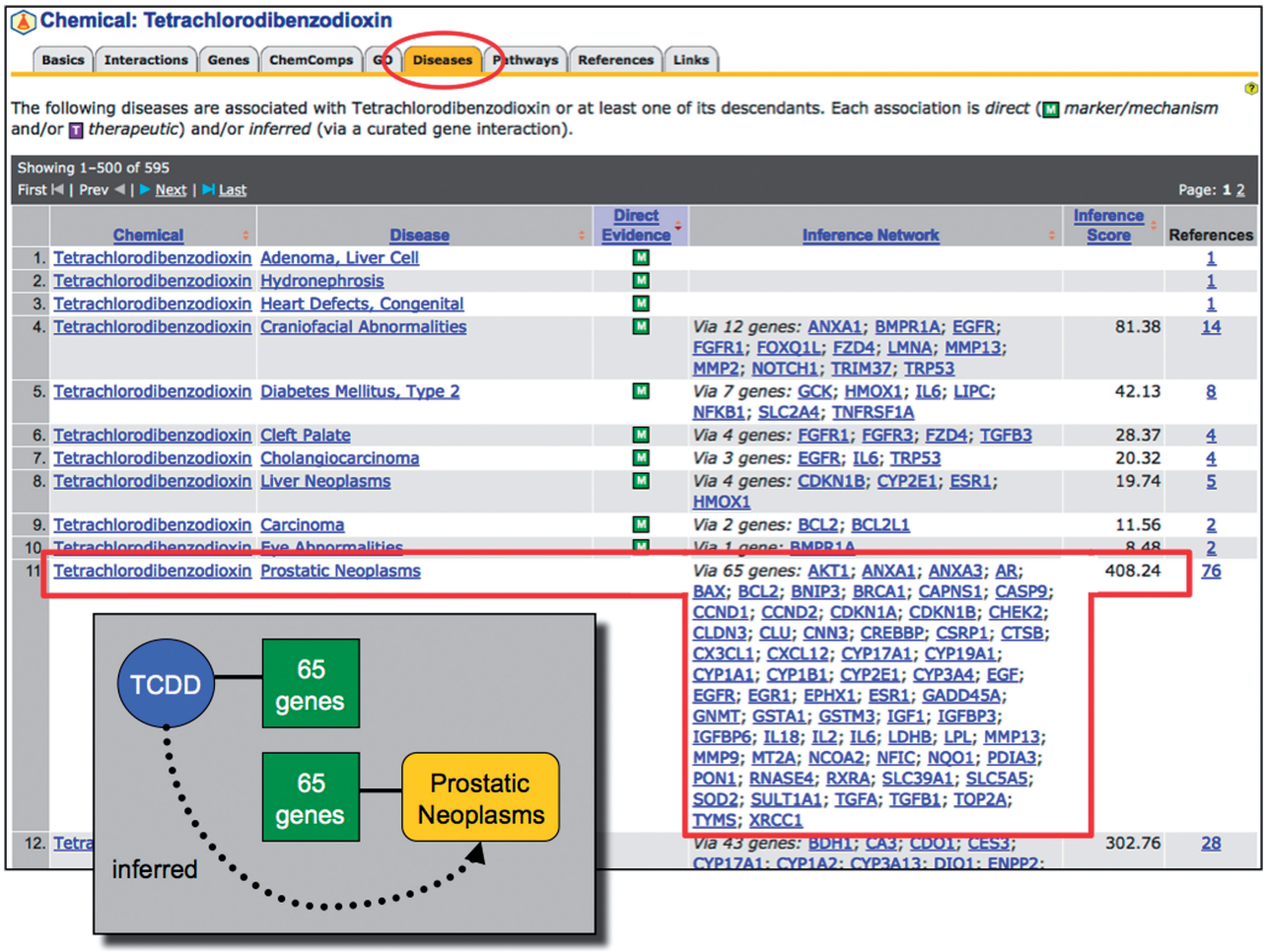


**Figure 2.** Network scores help rank chemical–disease inferences. On the highlighted 'Disease' data tab (red circle) for the chemical tetrachlorodibenzodioxin (TCDD), users find both direct and inferred disease associations. Inferred chemical–disease relationships are enhanced with inference network scores, allowing predicted relationships to be sorted and prioritized. In this partial screenshot, the highest inference score (408.24) associates TCDD with prostrate cancer based upon 65 genes that interact with TCDD and independently have a direct relationship with the disease (inset and red box). For more information about inference scores, click the 'Help' popup box (question mark link) in the upper-right corner of the text section.

what new chemicals, genes and diseases were curated for that month. Users can also request curation of specific papers or underrepresented chemicals using the 'Contact us' tab.

Lastly, we have started engaging the scientific community in reviewing the curation at CTD. When available, the email address of the corresponding author from a curated paper is captured by biocurators. After each monthly update, emails are automatically sent to the authors to alert them that their work has been curated and to ask them to review the data. To date, over 3900 authors have been notified for more than 4600 papers. This interaction with and feedback from the research community helps to ensure the high quality of curated data, and introduces CTD to potential new users.

## SUMMARY AND FUTURE DIRECTIONS

CTD provides detailed information about manually curated chemical–gene interactions, chemical–disease relationships and gene–disease relationships. By integrating these core data with other datasets, CTD helps turn knowledge into discoveries by identifying novel connections between chemicals, genes, diseases, pathways and GO annotations that might not otherwise be apparent using other biological resources.

Here we have highlighted the recent improvements to CTD, including new data content, expanded data sources, enhanced data features and new analytical tools that allow users to perform meta-analyses of the datasets. Users can also search with our query pages that allow a multitude of parameters to be queried simultaneously (e.g. GO annotations, pathway terms, chemical classes, types of interactions and diseases) to ask sophisticated questions such as: which transcription factors have their activity affected by heavy metals, or how does the chemical resveratrol affect the gene p53? Examples of how to run these queries and retrieve the results are provided in the FAQ section of the 'Help' menu tab.

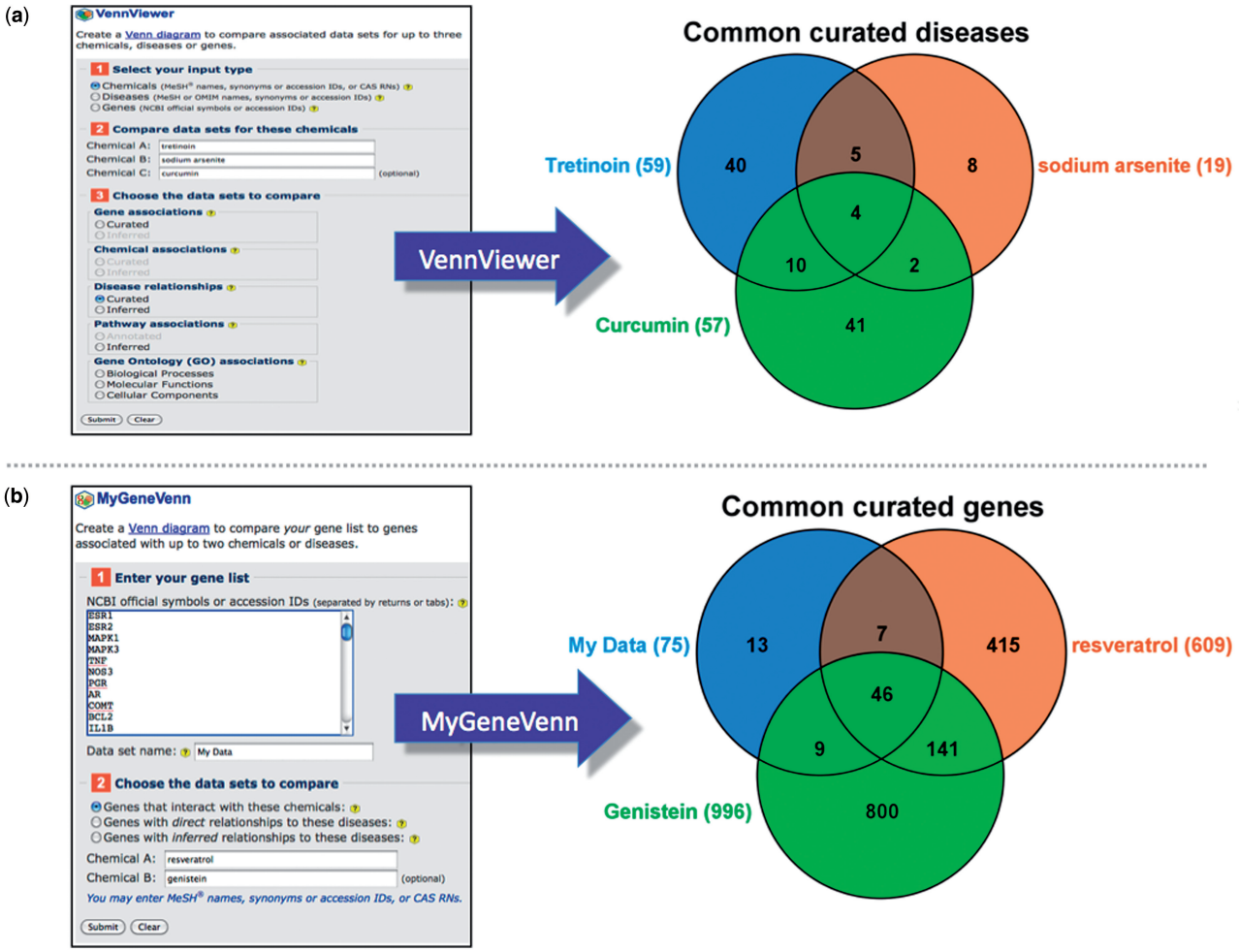


**Figure 3.** CTD provides new analytical tools. (**a**) VennViewer allows users to create online Venn diagrams to compare associated datasets for any three chemicals, genes or disease. Here, the three chemicals tretinoin, sodium arsenite and curcumin are analyzed to find common curated diseases, including four diseases shared by all three chemicals. (**b**) MyGeneVenn lets users upload gene lists of interest to find supporting data or to determine which genes from their set have an association with specified chemicals or diseases in CTD. Shown here, from a user's input of 75 genes ('My Data'), 46 genes are also found to interact with resveratrol and genistein, seven interact with resveratrol only and nine interact with genistein only.

In the future, we hope to expand the depth and breadth of the manually curated core data in CTD, include new inference scores for gene–disease predictions (to complement the inference scores for chemical–disease predictions) and provide a new metric ('DiseaseComps') to describe diseases related to each other based upon shared toxicogenomic profiles. As well, we plan to increase the utility of our inferred data for GO terms and pathways.

All of these features continue to make CTD a unique scientific resource that promotes understanding about the effects of environmental chemicals on human health and for generating testable hypotheses about the mechanisms underlying the etiology of environmental diseases.

## FUNDING

National Institutes of Health grants, National Institute of Environmental Health Sciences and the National Library of Medicine (R01 ES014065 and R01 ES014065-04S1 to CTD); INBRE program of the National Center for Research Resources (P20 RR016463). Funding for open access charge: NIEHS grant (R01 ES014065).

*Conflict of interest statement*. None declared.

## REFERENCES

1. McHale,C.M., Zhang,L., Hubbard,A.E. and Smith,M.T. (2010) Toxicogenomic profiling of chemically exposed humans in risk assessment. *Mutat. Res.*, [Epub ahead of print, 9 April].
2. 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.
3. Mattingly,C.J., Rosenstein,M.C., Colby,G.T., Forrest,J.N. Jr 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.
4. Mattingly,C.J., Rosenstein,M.C., Davis,A.P., Colby,G.T., Forrest,J.N. Jr and Boyer,J.L. (2006) The comparative

toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.*, **92**, 587–595.
5. Mattingly,C.J. (2009) Chemical databases for environmental health and clinical research. *Toxicol. Lett.*, **186**, 62–65.
6. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
7. Hernandez-Boussard,T., Whirl-Carrillo,M., Hebert,J.M., Gong,L., Owen,R., Gong,M., Gor,W., Liu,F., Truong,C., Whaley,R. *et al.* (2008) The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.*, **36**, D913–D918.
8. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
9. Lim,E., Pon,A., Djoumbou,Y., Knox,C., Shrivastava,S., Guo,A.C., Neveu,V. and Wishart,D.S. (2010) T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res.*, **38**, D781–D786.
10. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
11. Tomasulo,P. (2002) ChemIDplus-super source for chemical and drug information. *Med. Ref. Serv. Q.*, **21**, 53–59.
12. Kuhn,M., von Mering,C., Campillos,M., Jensen,L.J. and Bork,P. (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.
13. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**, W305–W311.
14. Sardana,D., Vasa,S., Vepachedu,N., Chen,J., Gudivada,R.C., Aronow,B.J. and Jegga,A.G. (2010) PhenoHM: human-mouse comparative phenome-genome server. *Nucleic Acids Res.*, **38 (Suppl.)**, W165–W174.
15. Chen,B., Dong,X., Jiao,D., Wang,H., Zhu,Q., Ding,Y. and Wild,D.J. (2010) Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinform.*, **11**, 255.
16. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
17. Glez-Pena,D., Gomez-Lopez,G., Pisano,D.G. and Fdez-Riverola,F. (2009) WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis. *Nucleic Acids Res.*, **37**, W329–W334.
18. Audouze,K., Juncker,A.S., Roque,F.J., Krysiak-Baltyn,K., Weinhold,N., Taboureau,O., Jensen,T.S. and Brunak,S. (2010) Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks. *PLoS Comput. Biol.*, **6**, e1000788.
19. 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.
20. Patel,C.J. and Butte,A.J. (2010) Predicting environmental chemical factors associated with disease-related gene expression data. *BMC Med. Genomics*, **3**, 17.
21. Wexler,P. (2004) The U.S. National Library of Medicines Toxicology and Environmental Health Information Program. *Toxicology*, **198**, 161–168.
22. 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.
23. Wiegers,T.C., Davis,A.P., Cohen,K.B., Hirschman,L. and Mattingly,C.J. (2009) Using text mining to enhance manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). *BMC Bioinform.*, **10**, 326.
24. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
25. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
26. Wishart,D.S., Knox,C., Guo,A.C., Cheng,D., Shrivastava,S., Tzur,D., Gautam,B. and Hassanali,M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.
27. 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.
28. 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. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
29. Chamie,K., DeVere White,R.W., Lee,D., Ok,J.H. and Ellison,L.M. (2008) Agent Orange exposure, Vietnam War veterans, and the risk of prostate cancer. *Cancer*, **113**, 2464–2470.