Bioinformatics, 34(18), 2018, 3235–3237
doi: 10.1093/bioinformatics/bty326
Advance Access Publication Date: 24 April 2018
Applications Note

OXFORD

Databases and ontologies

CTDquerier: a bioconductor R package for Comparative Toxicogenomics Database™ data extraction, visualization and enrichment of environmental and toxicological studies

Carles Hernandez-Ferrer[1,2,3,4] and Juan R. Gonzalez[1,2,3,*]

[1]Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain and [4]Computational Health Informatics Program, Boston Children's Hospital, Boston, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on July 4, 2017; revised on April 6, 2018; editorial decision on April 18, 2018; accepted on April 20, 2018

Abstract

Summary: Biomedical studies currently include a large volume of genomic and environmental factors for studying the etiology of human diseases. R/Bioconductor projects provide several tools for performing enrichment analysis at gene-pathway level, allowing researchers to develop novel hypotheses. However, there is a need to perform similar analyses at the chemicals-genes or chemicals-diseases levels to provide complementary knowledge of the causal path between chemicals and diseases. While the Comparative Toxicogenomics Database™ (*CTD*) provides information about these relationships, there is no software for integrating it into R/Bioconductor analysis pipelines. `CTDquerier` helps users to easily download *CTD* data and integrate it in the R/Bioconductor framework. The package also contains functions for visualizing *CTD* data and performing enrichment analyses. We illustrate how to use the package with a real data analysis of asthma-related genes. `CTDquerier` is a flexible and easy-to-use Bioconductor package that provides novel hypothesis about the relationships between chemicals and diseases.

Availability and implementation: `CTDquerier` R package is available through Bioconductor and its development version at https://github.com/isglobal-brge/CTDquerier.

Contact: juanr.gonzalez@isglobal.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomics has dramatically improved our understanding of the molecular origins of complex human diseases. However, the environment also has a strong influence on those diseases. Chemicals such as heavy metals, pesticides, solvents, paints, detergents, kerosene, carbon monoxide and drugs lead to unintentional poisonings at home, causing 193 000 deaths annually (http://www.who.int/quantifying_ehimpacts). Environmental factors such as air pollution and second-hand smoke are risk factors for adverse pregnancy outcomes such as low birth weight, prematurity and stillbirths (Khader *et al.*, 2011). Air pollution also exacerbates and increases asthma-related hospital admissions (MacIntyre *et al.*, 2013). There is a wealth of data that can provide biological insights into how environmental exposures affect human health. However, there is a shortage of tools for integrating these data to develop new hypotheses about the mechanisms underlying complex diseases.

The Comparative Toxicogenomics Database™(*CTD*) is a public resource for toxicogenomic information manually curated from peer-reviewed scientific literature, providing critical information about the interactions between environmental chemicals and gene products and their effect on human disease (Mattingly *et al.*, 2003).

CTD includes information on a triad of core interactions describing chemical–gene, chemical–disease and gene–disease relationships. An updated version contains more than 30.5 million toxicogenomic connections relating chemicals/drugs, genes/proteins, diseases, exposures, Gene Ontology (GO) annotations, pathways (KEGG/Reactome), and gene interaction modules (Davis *et al.*, 2017).

CTD integrates manually curated data reported in peer-reviewed literature. Inferred associations are established via curated chemical–gene interactions. These associations are putative molecular links between otherwise disconnected data that could be used to test hypotheses (Davis *et al.*, 2015).

The R/Bioconductor project provides several tools for performing enrichment analysis at the gene-pathway level. However, there is a lack of tools for providing this type of information at the chemicals-genes or chemicals-diseases levels. We have created the `CTDquerier` R/Bioconductor package to fill this gap. It facilitates the inclusion of *CTD* data in downstream statistical analyses in R/Bioconductor pipelines. Queries can be performed directly from R at the gene, chemical or disease levels. The package also includes a series of plots for visualizing results retrieved from *CTD*, and functions for performing enrichment analyses that can help in biological interpretation or for generating novel hypotheses.

2 The `CTDquerier` R/Bioconductor package

The `CTDquerier` R/Bioconductor package allows users to query *CTD* by gene (`query_ctd_gene` function), by chemical (`query_ctd_chem` function) and/or by disease (`query_ctd_dise` function) using single or multiple terms (Fig. 1A). These functions are described in the package's main vignette available in Bioconductor (http://bioconductor.org/packages/devel/bioc/vignettes/CTDquerier/inst/doc/vignette.html). The terms given to these functions are validated against the *CTD* vocabulary files that are retrieved from *CTD* and stored as local cache files using `BiocFileCache` R package (http://bioconductor.org/packages/BiocFileCache/). The terms that are validated are used by `CTDquerier` to perform HTTP/S request to *CTD* downloading the multiple results as TSV files. These TSV files are read as `DataFrames` (http://bioconductor.org/packages/S4Vectors/), and once the query is completed they are encapsulated in an S4 object of class `CTDdata`. Figure 1B depicts the type of information returned by each type of query. In all cases, a table containing all the information from *CTD* is obtained, while heatmaps, networks or barplots are created, depending on the type of query and the result the user wants to visualize.

2.1 CTDdata functionality

The three main functions included into `CTDquerier` return an object of class `CTDdata`. This is the main object of `CTDquerier` and encapsulates the data retrieved from *CTD* and ensures compatibility with other R/Bioconductor packages. These tables are stored as `DataFrames`. The information included in each table depends on the type of query performed (gene, chemical or disease query).

Three methods are provided for `CTDdata` objects: (i) `get_terms` retrieves the terms that are validated in *CTD* vocabulary files; (ii) `get_table` fetches data from *CTD* as an object of class `DataFrame` that can be used in third party packages; (iii) `enrich` performs a Fisher's exact test between two `CTDdata` objects. By default, this function uses the genes available in *CTD* as the gene universe. However, the user may indicate the gene universe using the argument gene_univ. `CTDquerier` includes a gene universe

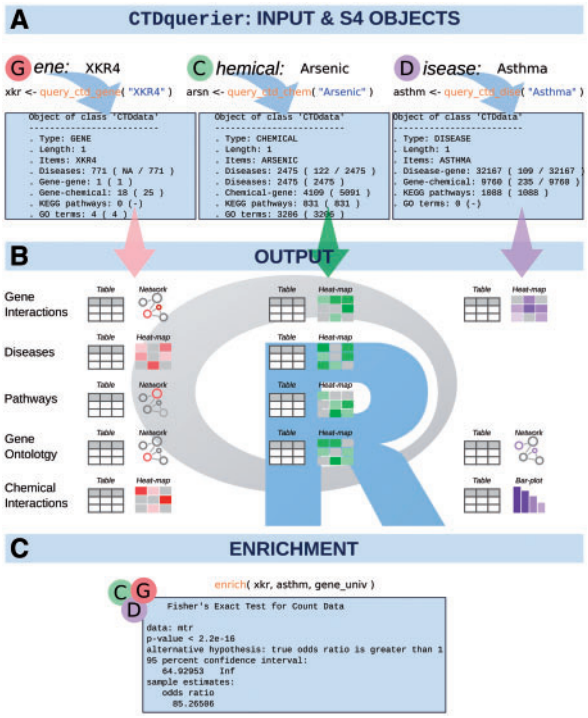

Fig. 1. (**A**) Illustrates the different *CTD* queries that can be performed using genes, chemicals and diseases (**B**) indicates the data retrieved into R and the possible visualization plots and (**C**) illustrates an example of how to perform enrichment analysis

obtained from HGNC (HUGO Gene Nomenclature Committee) (Supplementary Material Section 2.8).

2.2 CTDdata visualization

Figure 1B illustrates the plots that can be created using `CTDdata` objects (heatmaps, networks and barplots). The figures are generated using the R standard method `plot`. `CTDquerier` also provides different types of representations, depending on the type of query to *CTD*, as different sets of information may be available.

3 Results

3.1 Case study on environmental chemicals and asthma-related genes

We apply `CTDquerier` in a real data analysis. The Genetics of Asthma in Latino Americans (GALA) study aimed to identify novel asthma-associated loci in Latino populations using case-control admixture mapping (Torgerson *et al.*, 2012). The authors found a total of 305 asthma-related genes (i.e. GALA genes). Section 2 of Supplementary Material describes the R code used to address different scientific questions such as how many diseases are related to GALA genes, or which chemicals are associated with GALA genes. Other questions related to enrichment of GALA genes in gene sets related to a given disease or chemical are also answered.

Forty-eigth of the 305 GALA genes are not present in *CTD* (Supplementary Fig. S1). Table 1 depicts how many relationships with GALA genes are present in *CTD* at different levels and how many of them are curated relationships. For instance, there are 234 026 relationships between GALA genes and diseases, of which 2908 are curated. Of these, 303 unique diseases are linked to GALA

Table 1. Information obtained from the CTDdata created by querying the GALA genes in *CTD* (date: March 13, 2018)

	Curated	Total
Disease	2908	234 026
Gene–gene interactions	8925	11 599
Gene–chemical interactions	10 868	17 914
Pathways	1340	1340
GO terms	4715	4727

genes, supported by 441 curated relationships (Supplementary Material Section 2.5.2).

CTD also provides a statistic (Inference Score) for prioritizing inferences for hypothesis development, based on the topology of networks. The higher the score, the more likely the inference network has atypical connectivity (King *et al.*, 2012). Supplementary Figure S2 shows that the genes *IL12B*, *EGR1*, *EGFR*, *NGF* and *ITGAM* have the highest inference score. The high connectivity of these genes reflects their roles in asthma processes. The same information can be obtained at the chemicals-asthma level. Supplementary Figure S4 indicates that ozone and particulate matter are in the top-10 positions. In terms of the enrichment analysis, we observe that GALA genes are significantly enriched (OR for enrichment = 1.82, $P = 0.0266$) in the set of genes linked to air pollutant chemicals (Supplementary Material Section 2.8.2).

3.2 Comparing CTDquerier with *CTD*'s tools

Batch Query is a web-based tool provided by *CTD* that allows the user to query genes, diseases or chemicals from CSV-like files. Chapter 3 of the Supplementary Material includes six different comparisons between CTDquerier and Batch Query. As expected, the results obtained from both tools are identical.

Set Analyzer is another tool provided by *CTD* that allows the user to perform enrichment analyses. In chapter 4 of the Supplementary Material, we compare CTDquerier with this tool. Set Analyzer shows that GALA genes are enriched in gene sets related to neoplasm and the nervous system, among others. The Fisher exact test peformed using CTDquerier gives similar p-values for enrichment for those two diseases, 2.7×10^{-7} and 4.8×10^{-7}, respectively. Therefore, the results provided by CTDquerier and Set Analyzer are in agreement.

4 Conclusion

CTDquerier is a new R/Bioconductor package for retrieving, visualizing and performing enrichment analysis with data from *CTD*. The package can be integrated into pipelines designed to provide biological insights in a wide range of settings, such as genetic, toxicological and environmental studies that use standard R/Bioconductor tools to perform association analyses. The package includes functions for performing enrichment analysis at the gene level between different *CTD*'s queries. We illustrate the utility of CTDquerier by performing a real data analysis in genes obtained from a study of asthma. The validity of the results provided by CTDquerier has been tested by comparing them with results generated by the web-based tools provided by *CTD*.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness [MTM2015-68140-R] and the HELIX Project supported by European Commission FP7 [GA#308333]. C.H.F. is supported by funding from the National Institutes of Mental Health [NIMH R01MH107205]. ISGlobal is a member of the CERCA Programme, Generalitat de Catalunya.

Conflict of Interest: none declared.

References

Davis,A.P. *et al.* (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914–D920.

Davis,A.P. *et al.* (2017) The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.*, **45**, D972–D978.

Khader,Y.S. *et al.* (2011) The association between second hand smoke and low birth weight and preterm delivery. *Matern. Child Health J.*, **15**, 453–459.

King,B.L. *et al.* (2012) Ranking transitive chemical-disease inferences using local network topology in the comparative toxicogenomics database. *PLoS One*, **7**, e46524.

MacIntyre,E.A. *et al.* (2013) Traffic, asthma and genetics: combining international birth cohort data to examine genetics as a mediator of traffic-related air pollution's impact on childhood asthma. *Eur. J. Epidemiol.*, **28**, 597–606.

Mattingly,C.J. *et al.* (2003) The Comparative Toxicogenomics Database (CTD). *Environ. Health Perspect.*, **111**, 793–795.

Torgerson,D. *et al.* (2012) Case-control admixture mapping in latino populations enriches for known asthma-associated genes. *J. Allergy Clin. Immunol.*, **130**, 76–82.