Bioinformatics Advance Access published November 11, 2015

*Application notes*

# WormExp: a web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis

Wentao Yang[1], Katja Dierking[1] and Hinrich Schulenburg[1,*]

[1]Evolutionary Ecology and Genetics, Zoological Institute, CAU Kiel, Am Botanischen Garten 9, 24118 Kiel, Germany.

Associate Editor: Dr. Ziv Bar-Joseph

**ABSTRACT**

**Motivation:** A particular challenge of the current omics age is to make sense of the inferred differential expression of genes and proteins. The most common approach is to perform a gene ontology (GO) enrichment analysis, thereby relying on a database that has been extracted from a variety of organisms and that can therefore only yield reliable information on evolutionary conserved functions.

**Results:** We here present a web-based application for a taxon-specific gene set exploration and enrichment analysis, which is expected to yield novel functional insights into newly determined gene sets. The approach is based on the complete collection of curated high-throughput gene expression data sets for the model nematode *Caenorhabditis elegans*, including 1786 gene sets from more than 350 studies.

**Availability and implementation:** WormExp is available at http://wormexp.zoologie.uni-kiel.de.

**Contacts:** wyang{at}zoologie.uni-kiel.de, kdierking{at}zoologie.uni-kiel.de, or hschulenburg{at}zoologie.uni-kiel.de

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

## 1 INTRODUCTION

High-throughput molecular technologies have greatly enhanced our understanding of biological processes by characterizing expression changes of genes (microarray and RNA-Seq data) and proteins (proteomics data) or transcription factor targets and epigenetics states (ChIP-chip and ChIP-Seq data). These technologies usually yield hundreds or thousands of differentially regulated genes or proteins that are not always easy to interpret. Validation of the numerous differentially expressed genes is usually not possible. Uncovering the underlying organizational principles from such large gene lists requires computational and statistical approaches as well as precise biological reference information.

Gene set enrichment analysis represents a powerful tool to link the identified differentially expressed gene lists to biological processes and functions. They are based on the statistical evaluation of the overlap between the generated gene set and a specified reference list of genes. These enrichment analyses are usually based on public databases such as those defined by Gene ontology (Ashburner, et al., 2000) and KEGG pathways (Kanehisa and Goto, 2000). However, these existing databases have important drawbacks. First, the annotations are incomplete and only a subset of known genes are functionally annotated (King, et al., 2003). For example, functional information is only available for approximately 60% of the gene repertoire of the nematode *Caenorhabditis elegans*, one of the most intensively studied model organisms in biological research (Petersen, et al., 2015). Second, the included functional information is often imprecise, as it usually represents an extrapolation from experimental data of a different taxon and thus assumes a high level of functional conservation across evolution, which may not always be the case (Khatri and Drăghici, 2005). Third, functional information is predicted for most organisms from protein domains. Taxon-specific genes or protein domains may thus be missed. Taxon-specific gene sets, which explicitly consider taxon-restricted genes and also taxon-specific expression responses, are thus required for improved functional genomic analyses. Several applications such as GSEA (Subramanian, et al., 2005) and EASE (Hosack, et al., 2003) have been developed to permit performance of enrichment analyses with curated gene sets, derived for example from published expression studies in the same organism. Yet, a systematic assessment of the value of taxon-specific enrichment analyses is still missing. WormExp provides a web resource to explore such an approach for the nematode *Caenorhabditis elegans*.

This nematode has a well annotated genome sequence and is widely used as a powerful model organism in biological research. Over the last decade, more than 350 high-throughput expression studies have been published, covering a large variety of research themes, such as immunity, aging, development, and stress responses. The resulting lists of differentially expressed genes are publicly available and can be related to a specific experimental design, environmental condition, and/or gene defect. Because they capture a variety of inducible expression responses of this particular organism, they might be highly useful in interpreting new *C. elegans* gene lists (Engelmann, et al., 2011; Yang, et al., 2015) or predicting candidates for downstream analysis (Block, et al., 2015). In this manuscript, we present WormExp, a web-based application for gene set enrichment analysis in *C. elegans*. We collated nearly all published high-throughput expression data sets of *C. elegans* from public databases and also the available literature. We classified these gene sets in nine categories according to the experimental designs or specific condition used. WormExp accepts Wormbase (Harris, et al., 2010) identifiers (IDs), sequence names, gene names or a mix-

*To whom correspondence should be addressed.

ture of these as input. It offers tools for performing an enrichment analysis based on the taxon-specific 1786 gene sets, searching specific gene lists, and downloading complete data sets.

## 2 METHODS AND FEATURES

WormExp utilizes a curated database built from published high-throughput expression studies in the nematode *C. elegans*. The database can be downloaded and will be updated continuously. Users start their analysis by uploading a gene list, for example a set of genes, whose expression is induced upon *C. elegans* exposure to a certain condition. The user then has two options (Fig. 1 and workflow in manual): (i) perform an enrichment analysis using either the entire database as reference or selected categories of gene lists (e.g. “mutants”), or (ii) search for overlaps between the uploaded gene set and specific gene lists (e.g. “up *pmk-1* mutant”), selected from the database with the help of keywords (e.g. “pmk-1”). For enrichment analysis, WormExp employs the adjusted Fisher exact test from the program EASE, which penalizes or removes one gene within a given gene set from the test list and calculates the p-value (for further details, see manual in supplementary file). This modulation makes the Fisher exact test more robust when applied on gene sets supported by few genes, thus reducing false positives (for details, please see (Hosack, et al., 2003)). WormExp is available as a webserver with an interface from InterMine (Smith, et al., 2012), developed by Java 2 Enterprise System (J2EE) and Java Remote Method Invocations (RMI). RMI ensures fast responses due to memory-oriented query. See supplementary files 1 and 2 for manual and detailed information.

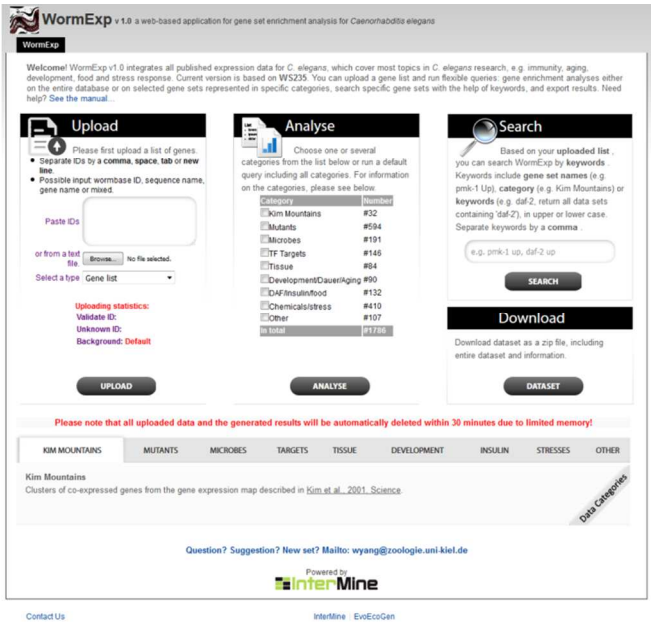


**Fig.1 Homepage of WormExp**

### 2.1 Data Structure

Currently, the WormExp database includes 1786 gene sets derived from 361 studies and collated from the NCBI GEO database (Barrett, et al., 2009), ArrayExpress database (Brazma, et al., 2003), Stanford microarray database (Sherlock, et al., 2001), Princeton University MicroArray database, and directly from the literature (Supplementary Data 1). According to experimental design and used conditions, we classified these 1786 gene sets into nine categories: Kim Mountains ((Kim, et al., 2001)); Mutants (differentially expressed genes in mutants or upon RNA interference-silencing of a particular gene); Microbes (exposure to various microorganisms), TF Targets (transcription factor targets inferred by knock-down/knock-out of the respective transcription factors), Tissue (tissue specific expression), Development/Dauer/Aging (differential expression in the various developmental stages and during aging), DAF/Insulin/food (differential expression in response to food, starvation, or insulin-like receptor activation/de-activation), Chemicals/stress (exposure to chemical compounds or other stressors), and Other (all gene sets not included above). The database will be updated regularly to integrate new *C. elegans* expression studies.


## ACKNOWLEDGEMENTS

We thank the Schulenburg lab for advice and the Kiel University computer center, especially S. Lorenz and U. Schwarz, for support.

*Funding*: The work was funded by grants of the German Science foundation to HS (DFG grants SCHU 1415/8 and SCHU 1415/9). WY is additionally supported by the International Max-Planck Research School for Evolutionary Biology.


## REFERENCES


Ashburner, M., et al. (2000) Gene Ontology: tool for the unification of biology, Nature genetics, 25, 25-29.

Barrett, T., et al. (2009) NCBI GEO: archive for high-throughput functional genomic data, Nucleic acids research, 37, D885-D890.

Block, D.H., et al. (2015) The Developmental Intestinal Regulator ELT-2 Controls p38-Dependent Immune Responses in Adult C. elegans.

Brazma, A., et al. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI, Nucleic acids research, 31, 68-71.

Engelmann, I., et al. (2011) A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in C. elegans, PloS one, 6, e19055.

Harris, T.W., et al. (2010) WormBase: a comprehensive resource for nematode research, Nucleic acids research, 38, D463-D467.

Hosack, D.A., et al. (2003) Identifying biological themes within lists of genes with EASE, Genome Biol, 4, R70.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, Nucleic acids research, 28, 27-30.

Khatri, P. and Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems, Bioinformatics, 21, 3587-3595.

Kim, S.K., et al. (2001) A gene expression map for Caenorhabditis elegans, Science, 293, 2087-2092.

King, O.D., et al. (2003) Predicting gene function from patterns of annotation, Genome research, 13, 896-904.

Petersen, C., Dirksen, P. and Schulenburg, H. (2015) Why we need more ecology for genetic models such as C. elegans, Trends in Genetics, 31, 120-127.

Sherlock, G., et al. (2001) The stanford microarray database, Nucleic Acids Research, 29, 152-155.

Smith, R.N., et al. (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data, Bioinformatics, 28, 3163-3165.

Subramanian, A., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences of the United States of America, 102, 15545-15550.

Yang, W., et al. (2015) Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode Caenorhabditis elegans toward pathogenic Bacillus thuringiensis, Developmental & Comparative Immunology, 51, 1-9.