**Table of Contents:**

[1 Goal of the documentation 2](#_Toc94180219)

[2 Scope and Responsibilities 2](#_Toc94180220)

[3 Term/Definition/Abbreviation 2](#_Toc94180221)

[4 Procedure/ Workflow 2](#_Toc94180222)

[4.1 Software 2](#_Toc94180223)

[4.2 Security details 2](#_Toc94180224)

[4.3 Method procedure 3](#_Toc94180225)

[4.3.1 General Notes 3](#_Toc94180226)

[4.3.2 Downloading WormSource 3](#_Toc94180227)

[4.3.3 Finding new GEO datasets 4](#_Toc94180228)

[4.3.4 Transcriptomics File 5](#_Toc94180229)

[4.3.5 Assembling dataset 6](#_Toc94180230)

[4.3.6 Manual Data Analysis 7](#_Toc94180231)

[4.3.7 WormBase ID 7](#_Toc94180232)

[4.3.8 Fusion of Files 7](#_Toc94180233)

[4.3.9 TomCat pilot run 7](#_Toc94180234)

[4.3.10 Updating database 7](#_Toc94180235)

[5 Document History 7](#_Toc94180236)

[6 Publication bibliography 7](#_Toc94180237)

**Further documents**:

(evogen 2015) Manual for WormExp

(Yang et al. 2016) WormExp :WormExp: a web-based application for a Caenorhabditis elegans-specific gene expression enrichment analysis

Website <https://wormexp.zoologie.uni-kiel.de/wormexp/>

(Dozmorov 2016) GEOparse documentation

datafinder jupyter notebook file

1. **Goal of the documentation**

This documentation describes the procedure to update the web-based application WormExp.

1. **Scope and Responsibilities**

This documentation is only valid for AG Schulenburg.

|  |  |
| --- | --- |
| **Function** | **Responsibilities** |
| Employee/User | Updating WormExp  Is responsible for the correct use and update of the application. |

1. **Term/Definition/Abbreviation**

|  |  |
| --- | --- |
| GEO | Gene Expression Omnibus |

1. **Procedure/** **Workflow**

## Software

|  |  |  |  |
| --- | --- | --- | --- |
| Software/Website | Specification | Version | Source/Link |
| WormExp | web-based application for a taxon-specific gene set exploration and enrichment analysis | WormExp v1.0 | https://wormexp.zoologie.uni-kiel.de/wormexp/ |
| Python | Programming language | 3.8.11 | https://www.python.org/ |
| Anaconda | Python package distribution and management | 2020.11 | https://www.anaconda.com/ |
| GEOparse | Python library to access Gene Expression Omnibus Database (GEO) | 2.0.3 | https://geoparse.readthedocs.io/en/latest/GEOparse.html |
| Jupyter Notebook | Web-based environment for working with notebooks | 6.4.0 | https://jupyter-notebook.readthedocs.io/en/stable/index.html |
| Matplotlib | Python data visualization tool | 3.4.2 | https://matplotlib.org/ |
| Numpy | Core package for scientific computing with Python. | 1.20.3 | https://numpy.org/ |
| Pandas | Library for tabular data structures | 1.3.2 | https://pandas.pydata.org/docs/index.html# |
| Biopython | Library for biological computation written in Python | 1.78 | https://biopython.org/ |

## Security details

N/A

## Method procedure

### General Notes

This documentation describes the general procedure on how to update the database behind WormExp. The details in the procedure can be modified depending on the user’s operation system, experience, etc. The scripts and procedures used here were based on a Python (version 3.8.11) script in a virtual environment managed by anaconda (version 2020.11). The virtual environment contained several additional libraries: GEOparse (version 2.0.3), Jupyter (version 6.4.0), Matplotlib (version 3.4.2), Numpy (version 1.20.3), Pandas (version 1.3.2) and Biopython (version 1.78).

These libraries are essential for a correct function of the provided script to find new GEO datasets uploaded into Pubmed GEO. However, GEOparse is also available for R and the whole procedure can therefore be transferred to R if wished.

The data should be collected in a separate folder and new Excel files and only merged with the existing database at the very end. Additionally, a pilot run should be included to make sure that everything works appropriately

### Downloading WormSource

The complete database source code and background information can be downloaded on <https://wormexp.zoologie.uni-kiel.de/wormexp/> in the category “Download” and “Dataset”.

Graphical user interface

Description automatically generated

Figure 1: Download of WormSource

The file contains a folder “WormSource”, which holds the complete database information. It contains another folder WormExp, which holds the java source files to run it. Furthermore, it contains several .txt files:

|  |  |  |
| --- | --- | --- |
| File | Description | needs Update? |
| c\_elegans.WS283.geneIDs | contains Wormbase ID, gene name and gene ID for every gene | yes |
| reference | contains dataset name and link to publication | yes |
| Chemicalexposure-otherStress | contains datasets categorized to Chemicals/Stress | yes |
| DAF Insulin food | contains datasets categorized to DAF/Insulin/food | yes |
| Development-Dauer-Aging | contains datasets categorized to Development/Dauer/Aging | yes |
| Kim Mounts | contains datasets categorized to Kim Mountains | yes |
| Mutants | contains datasets categorized to Mutants | yes |
| Other | contains datasets categorized to Other | yes |
| Pathogen | contains datasets categorized to Microbes | yes |
| Targets | contains datasets categorized to TF Targets | yes |
| Tissue-specific | contains datasets categorized to Tissue | yes |
| Epigenetics | contains datasets categorized to Tissue | yes |
| WormExp\_info | has key information about the datasets, like number of genes, links to publications and methodology | yes |
| datafinder.ipynb | jupyter notebook that contains script and instruction to use GEOparse to find new datasets | only for improvements |
| Wormbase\_version\_changes | contains information about changes in WormBase IDs the last years | yes |

* Description of WormExp folder
* Add Documentation and datafinder.py ipynb
* Add WormExp\_info.csv without any fancy schnickschnak
* Keep worked over dataset file?
* Add pdf document of this documentation and project (?)
* For group: PowerPoint und Excel File mit bereits gefunden Datasets + Documentation

### Finding new GEO datasets

To find new GEO datasets, the jupyter notebook “datafinder” can be used. It contains scripts and instructions on how to use it. The script uses the API GEOparse and searches for datasets depending on the inserted query. During the start of this project an analysis was conducted to ensure the used query contains the correct datasets. The results are documented in the supporting pdf document and given to the project owner.

Adaptions to the query should only be made in respect to the publication date. As of 02/2022 the database contains datasets until 2018.

The script will create a separate Excel file called “GSE\_query\_results” in which further (manual) work will be conducted. This Excel file is the backbone of all further investigation and will be described in detail in the following section “Transcriptomics File”. It must be noted that in order to use the datafinder notebook, the software from section 4.1 has to be installed.

### Transcriptomics File

The datafinder script produces an Excel file that contains detailed information about every dataset found with GEOparse. It shows not only the exact title of the dataset, but also its geo\_accession number, publication date, contributors, etc. etc. See Figure 2 for an example.

Calendar

Description automatically generated with low confidence

Figure 2: Exempt from original Transcriptomics file

The main task here is to start sorting through the found datasets and finding out which datasets are useful for the updates. It has been decided to concentrate mostly on datasets that already possess an in-depth transcriptomics analysis done by the respective scientists. Datasets that only possess raw data can be ignored until otherwise stated. For a better overview, the transcriptomics file was transformed, and colors have been introduced (see Figure 3) to show which data has supplementary data available (blue), which datasets contain only raw data (yellow), and which datasets can be excluded (red). Exclusion of datasets was mainly due to no available paper or dataset was not focused on differential gene expression.

Table

Description automatically generated

Figure 3: Excempt from the Transcriptomics file, after initial preparation

Additionally, new columns have been introduced. One of the main tasks in updating the database is to sort found datasets to categories mentioned in 4.3.2. If a dataset can be sorted to more than one category, the respective row is copied and added directly underneath (see section 4.3.5 for more detailed information about categories). Furthermore, columns “Category\_check”, “Dataset\_check”, and “Comment” were added for supervision. Those categories will only be switched to “done” when categories and assembled dataset (see 4.3.6) were checked by supervisors and no problems occurred.

### Categorizing datasets

To-do

### Assembling dataset

Next to the transcriptomics file, another file must be updated for the database. For the following dataset assembly, the main work will be conducted in the WormExp\_info file. All new and added datasets must be added in the same manner as can be seen in the existing file. The columns are described in detail in the following table:

|  |  |
| --- | --- |
| Column | Description |
| WormBaseVersion | contains information which WormBase Version was used to map Entrez IDs to gene IDs. If n.a. no information was given. More information in 4.3.6. |
| Categories | same as in transcriptomics file. If more than one category applies, dataset will be added for every category. |
| Gene Set name | Explanation for the assembled gene set. |
| number\_genes | number of genes collected for respective gene set |
| Refs | Reference to paper |
| Data From | gives information where the gene set was found in the paper |
| Selection\_criteria | Shows which selection criteria was applied when extracting the data set. More information in 4.3.5.1 |
| decided\_by | gives information if the selection\_criteria was given by the authors or if it was decided by the assembler |
| Rawdata | Information to GSE accession number |
| Additional | Column for additional information |
| Comment | Column for comments |

Assembling the dataset is the trickiest and most error-prone part in this work. This work cannot be streamlined, as every scientist analyzed their work differently and uploaded it in different places, and in various formats. However, most of the times a differential gene expression analysis is uploaded in a separate table and can be found in the supplementary of the respective paper. Depending on the authors, selection criteria are more or less strict. The assembler must decide in most cases which genes to extract. In 4.3.5.1 some guidelines for selection criteria have been decided.

For every GSE number an overall gene set name should be chosen that describes sufficiently the experiment conducted. Every gene set extracted from the respective experiment should be saved in a .txt file that follows this name convention: [overall gene set name] \_[author]. In this .txt file all gene sets belonging to this experiment should be saved. Assembled gene sets should not be added yet to the category files in WormSource! This fusion step will follow after every gene set has been checked over by supervision and no further changes are made.

#### Criteria for supplementary data filtering

As mentioned, every scientist employed their own significance selection criteria. If possible, selection criteria from the authors should be used. These criteria will be added in the column „selection\_criteria” and “decided\_by”.

Exceptions come into play if the selection criteria are not strict enough (e.g., p-value > 0.1 without any adjustments and without corrections). In general, p-fdr/padj < 0.01 and log-foldchange >= 2 (or <= -2) should be used.

### WormBase ID

* <https://wormbase.org//tools/mine/simplemine.cgi>
* If Wormbase IDs were not found, delete them from data set
* If multiple entries, take both
* Adjust number of genes found in WormExp\_info
* Add wormbase version in WormExp\_info that was used for mapping
* Collect all data sets that have a Wormbase ID mapping (from the authors or yourself) in the respective categories in n separate text files from the current data base
* Describe procedure in simplemine (which boxes are checked)

### TomCat pilot run

* Use first new data sets for TomCat pilot test run
* Then fuse all category files for a final test

### Updating database

* Maybe change file “pathogen” to “microbes”?
* Added epigenetics category
* Other changes?
* How to update reference file

### Checklist

|  |  |
| --- | --- |
| **To-Do** | **Updated** |
| Wormbase\_version\_changes | yes |
| c.elegans.WS283.geneIDs | updated to from WS235 to WS283 |
| reference | Yes |
| Chemicalexposure-otherStress | Yes |
| DAF Insulin food | Yes |
| Development-Dauer-Aging | Yes |
| Kim Mounts | no |
| Mutants | Yes |
| Other | no |
| Pathogen (Microbes) | Yes |
| Targets | Yes |
| Tissue-specific | Yes |
| Epigenetics | Newly created; yes |
| WormExp\_info | fuse to one file |
| datafinder.ipynb | newly created; |

1. **Document History**

Reason for change

|  |  |  |
| --- | --- | --- |
| **Version number** | **Description of change** | **Valid from:** |
| 01 | 1. Version for establishing internal standards | 27.01.2022 |

Publication bibliography

Dozmorov, Mikhail (2016): GEOparse. Reading the NCBI’s GEO microarray SOFT files in R/BioConductor: public domain.

evogen (2015): Manual for WormExp, 2015. Available online at https://academic.oup.com/bioinformatics/article/32/6/943/1744078.

Yang, Wentao; Dierking, Katja; Schulenburg, Hinrich (2016): WormExp: a web-based application for a Caenorhabditis elegans-specific gene expression enrichnment analysis. In *Bioinformatics Advance Access* 32 (6).